The Telecommunications and Data Acquisition Progress Report 42-95

July–September 1988

E. C. Posner
Editor

November 15, 1988

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Telecommunications and Data Acquisition Progress Report 42-95

July–September 1988

E. C. Posner
Editor

November 15, 1988

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Preface

This quarterly publication provides archival reports on developments in programs managed by JPL's Office of Telecommunications and Data Acquisition (TDA). In space communications, radio navigation, radio science, and ground-based radio and radar astronomy, it reports on activities of the Deep Space Network (DSN) and its associated Ground Communications Facility (GCF) in planning, in supporting research and technology, in implementation, and in operations. Also included is TDA-funded activity at JPL on data and information systems and reimbursable DSN work performed for other space agencies through NASA. The preceding work is all performed for NASA's Office of Space Operations (OSO). The TDA Office also performs work funded by two other NASA program offices through and with the cooperation of the Office of Space Operations. These are the Orbital Debris Radar Program (with the Office of Space Station) and 21st Century Communication Studies (with the Office of Exploration).

In the search for extraterrestrial intelligence (SETI), the *TDA Progress Report* reports on implementation and operations for searching the microwave spectrum. In solar system radar, it reports on the uses of the Goldstone Solar System Radar for scientific exploration of the planets, their rings and satellites, asteroids, and comets. In radio astronomy, the areas of support include spectroscopy, very long baseline interferometry, and astrometry. These three programs are performed for NASA's Office of Space Science and Applications (OSSA), with support by the Office of Space Operations for the station support time.

Finally, tasks funded under the JPL Director's Discretionary Fund and the Caltech President's Fund which involve the TDA Office are included.

This and each succeeding issue of the *TDA Progress Report* will present material in some, but not necessarily all, of the following categories:

OSO Tasks:
- DSN Advanced Systems
  - Tracking and Ground-Based Navigation
  - Communications, Spacecraft-Ground
  - Station Control and System Technology
  - Network Data Processing and Productivity
- DSN Systems Implementation
  - Capabilities for Existing Projects
  - Capabilities for New Projects
  - New Initiatives
  - Network Upgrade and Sustaining
- DSN Operations
  - Network Operations and Operations Support
  - Mission Interface and Support
  - TDA Program Management and Analysis
- Communications Implementation and Operations
- Data and Information Systems
- Flight-Ground Advanced Engineering

OSO Cooperative Tasks:
- Orbital Debris Radar Program
- 21st Century Communication Studies

OSSA Tasks:
- Search for Extraterrestrial Intelligence
- Goldstone Solar System Radar
- Radio Astronomy

Discretionary Funded Tasks

Contents

OSO TASKS
DSN Advanced Systems
TRACKING AND GROUND-BASED NAVIGATION

COMMUNICATIONS, SPACECRAFT–GROUND

STATION CONTROL AND SYSTEM TECHNOLOGY

DSN Systems Implementation
CAPABILITIES FOR EXISTING PROJECTS

NETWORK UPGRADE AND SUSTAINING

Deriving a Geocentric Reference Frame for Satellite Positioning and Navigation

R. P. Malla†
S.-C. Wu
Tracking Systems and Application Section

With the advent of Earth-orbiting goedetic satellites, nongeocentric datums or reference frames have become things of the past. Accurate geocentric three-dimensional positioning is now possible and is of great importance for various geodetic and oceanographic applications. While relative positioning accuracy of a few centimeters has become a reality using very-long-baseline interferometry (VLBI), the uncertainty in the offset of the adopted coordinate system origin from the geocenter is still believed to be on the order of 1 meter. Satellite laser ranging (SLR), however, is capable of determining this offset to better than 10 cm, but this is possible only after years of measurements. Global Positioning System (GPS) measurements provide a powerful tool for an accurate determination of this origin offset. Two strategies are discussed in this article. The first strategy utilizes the precise relative positions that have been predetermined by VLBI to fix the frame orientation and the absolute scaling, while the offset from the geocenter is determined from GPS measurements. Three different cases are presented under this strategy. The reference frame thus adopted will be consistent with the VLBI coordinate system. The second strategy establishes a reference frame by holding only the longitude of one of the tracking sites fixed. The absolute scaling is determined by the adopted gravitational constant (GM) of Earth; and the latitude is inferred from the time signature of Earth rotation in the GPS measurements. The coordinate system thus defined will be a geocentric Earth-fixed coordinate system. A covariance analysis shows that geocentric positioning to an accuracy of a few centimeters can be achieved with just one day of precise GPS pseudorange and carrier phase data.

I. Introduction

The fully operational Global Positioning System (GPS) will consist of at least 18 satellites distributed in six orbital planes [1]. This system will allow a user, anywhere on the Earth or in a low Earth orbit, to view at least five satellites most of the time. Two precision data types can be derived from the GPS transmitted signals: P-code pseudorange and carrier phase at two L-band frequencies [2]. These precision data types provide the opportunity to produce geodetic measurements accurate to the centimeter level [3] and orbit determination of low Earth orbiters to the subdecimeter level [4]. The ephemerides for the GPS satellites, as distributed by Naval Surface Weapon Center (NSWC), are based upon the World Geodetic System

†Member of Professional Staff, Sterling Software, Pasadena, CA.

(WGS 84) [5], and their accuracy is on the order of 10 meters [6]. In applications where high precision is essential, the GPS satellite orbits need to be adjusted to a much higher precision, along with all the other parameters in the network [3], [4]. The GPS satellites can be simultaneously observed from several sites in a geodetic network. Within such a network a few fiducial tracking sites are included [7]. The relative positions of these fiducial sites are known to a higher level of precision, typically a few centimeters, as a result of repeated measurements of the baselines using very-long-baseline interferometry (VLBI) [8]. Based upon these highly precise relative positions of the fiducial sites, filter strategies can be designed to adjust the satellite orbits to enhance their accuracy to far better than 10 meters [9]. The ephemerides thus adjusted now refer to the same coordinate frame in which the fiducial baselines are known. It is generally believed that the best VLBI coordinate system origin approximates the geocenter to about 1 meter. The satellite laser ranging (SLR) technique is capable of realizing the geocenter offset to better than 10 cm, but this is possible only after years of observations.

Although absolute positioning is of less interest for geodynamic applications, it can be an important factor when tracking deep space vehicles, and it is essential for orbit determination of Earth-observing satellites, such as NASA's Ocean Topography Experiment (TOPEX), to be launched in late 1991 [10]. This article investigates two strategies for precise determination of the geocenter, thus establishing a geocentric coordinate frame for GPS measurements. In the first strategy, GPS P-code pseudorange and carrier phase measurements are made from a set of globally distributed tracking stations. A network consisting of six stations appears to be appropriate. Of these, three are the fiducial sites whose relative location has been well determined by VLBI. Since it is the relative location, rather than the absolute location, of the fiducial sites that is well determined by VLBI, only baseline coordinates should be fixed to define the orientation and absolute scaling of the reference frame. The geocenter position and the coordinates of other, nonfiducial sites are to be adjusted together with the GPS orbits. The coordinate frame thus defined is consistent with the VLBI frame, with improved geocenter offset. Three different cases are discussed under this strategy.

An alternate strategy is to simultaneously adjust the GPS orbits and geodetic station coordinates with respect to one reference site in the network whose longitude is held fixed. The absolute scaling is determined by the adopted gravitational constant GM of Earth; the station heights are inferred from the adjusted periods of GPS orbits and the pseudorange measurements; and the latitude is inferred from the time signature of Earth rotation in the GPS measurements. The coordinate system thus defined will be an Earth-centered, Earth-fixed (ECEF) coordinate frame. The solution is free from any a priori uncertainty of site positions, and the inferred reference frame is strictly self-contained. This type of technique has been adopted by the satellite laser ranging (SLR) and lunar laser ranging (LLR) communities [11]. The coordinate origin offset from the geocenter is given by the weighted mean coordinate offsets of all stations in the network.

A covariance analysis was performed estimating the accuracy with which the geocenter position can be determined using the two strategies. This analysis indicates that the geocenter position can be determined to an accuracy of a few centimeters with just one day of precision pseudorange and carrier phase data. Such precise knowledge of absolute position of the coordinate system origin is essential to the orbit determination of TOPEX, which requires an altitude accuracy of 13 cm or better.

II. Coordinate Reference Frame

A rectangular coordinate system is defined such that the Z axis coincides with the mean spin axis of the Earth as defined by the CIO pole; the X axis lies in the mean equatorial plane, which is perpendicular to the Z axis, and passes through the mean Greenwich astronomic meridian as defined by the BIH; the Y axis completes the right-handed Earth-fixed cartesian system. The origin of the coordinate system may be defined as the center of mass of the Earth. But the imperfect knowledge of the geocenter location limits the precise location of this origin.

Figure 1 gives the definition of the World Geodetic System 84 (WGS 84). The almanac and the ephemerides of GPS satellites are given in this coordinate system [6]. The coordinates of the ground stations derived by observing the GPS measurements will also be in the WGS 84 reference frame. But it should be noted that the absolute accuracy of any geocentric position determination depends upon the knowledge of the location of the geocenter relative to the assumed origin. The coordinate system thus defined is an ECEF coordinate system which rotates at a constant mean rate around a mean astronomic pole. Such a system is also called a conventional terrestrial system (CTS). However, events occur in an instantaneous real world, which is in a coordinate system different from the CTS. Therefore it is required to mathematically relate CTS to an instantaneous terrestrial system (ITS). This relationship is a transformation through a wobble [W] and a spin [S]:

$$\mathbf{X}_{ITS} = [S]\,[W]\,\mathbf{X}_{CTS} \tag{1}$$

where the **X**'s are position vectors. The wobble [W] is given by

$$[W] = \mathbf{R}_x(-y_p)\,\mathbf{R}_y(x_p) \tag{2}$$

where $\mathbf{R}_r(p)$ denotes a matrix of rotation, by an amount p about the r axis; x_p and y_p define the pole motion. The sign convention used is in accordance with the BIH convention. The spin is given by

$$[\mathbf{S}] = \mathbf{R}_z(-\text{GAST}) \tag{3}$$

where GAST is the Greenwich Apparent Sidereal Time given by

$$\text{GAST} = \text{GMST}^{0\text{h UT}} + \bar{\omega}\,(t_{df} + \text{UT1} - \text{UTC}) + \Delta\Psi \cos\epsilon \tag{4}$$

$\text{GMST}^{0\text{h UT}}$ is the Greenwich Mean Sidereal Time at 0 hour UT, which is obtained from Newcomb's equation adjusted with respect to J2000 [12], $\bar{\omega}$ is the mean rate of advance of the GMST per day, and t_{df} is the day fraction in UTC of time of observation. The last term in Eq. (4) is commonly known as the equation of the equinoxes, where $\Delta\Psi$ refers to the nutation in longitude and ϵ is the true obliquity of the ecliptic of date.

In general, celestial bodies are expressed in the Conventional Inertial System (CIS). Position vectors in this system can be transformed into ITS through a nutation [N] and a precession [P] [13]:

$$\mathbf{X}_{\text{CIS}} = [\mathbf{P}]\,[\mathbf{N}]\,\mathbf{X}_{\text{ITS}} \tag{5}$$

The nutation [N] is given by

$$[\mathbf{N}] = \mathbf{R}_x(-\epsilon_0)\,\mathbf{R}_z(-\Delta\Psi)\,\mathbf{R}_x(\epsilon_0 + \Delta\epsilon) \tag{6}$$

where ϵ_0 is the mean obliquity of date; the nutation angles $\Delta\Psi$ and $\Delta\epsilon$ are computed from IAU 1980 nutation series [12] expressed with respect to J2000. The precession [P] is given by

$$[\mathbf{P}] = \mathbf{R}_z(-\zeta_a)\,\mathbf{R}_y(-\theta_a)\,\mathbf{R}_z(z_a) \tag{7}$$

where ζ_a, θ_a, and z_a are the standard precession rotation angles [14]. Therefore, the position vectors in the reference frame WGS 84, which is one of the CTS, can be expressed with respect to the CIS using the above transformations.

The SLR system has matured enough to establish its own independent coordinate system. The dynamic technique used to establish such a system depends heavily upon a precise definition of the coordinate frame adopted by the tracking network. This includes the definition of polar motion and the Earth's fundamental constants, such as the gravitational constant (GM), the dynamical form factor (J_2), and the speed of light. Because satellite (LAGEOS, STARLETTE, etc.) position vectors are described in an inertial frame while ground station vectors are described in an ECEF frame, they need to be related by the above coordinate transformations. Processing of SLR long-arc data has been successful in simultaneously solving for station vectors, satellite orbits, and earth orientation parameters to precisions of few centimeters.

III. Strategies to Determine the Origin Offset from the Geocenter

For the past several years the fundamental concept behind accurate GPS orbital adjustment has been that of the fiducial network [7]. A fiducial network consists of three or more tracking stations whose (relative) positions have been determined in an Earth-fixed coordinate frame to a very high accuracy, usually by VLBI. Several receivers at other, less accurately known, stations also observe simultaneously the GPS satellites along with the fiducial network. The data are then brought together to simultaneously adjust the GPS satellite orbits and the positions of the nonfiducial sites. Thus the fiducial stations established by VLBI provide a self-consistent Earth-fixed coordinate system with respect to which the improved GPS satellite orbits and the nonfiducial stations can be expressed to a greater accuracy. At the same time the coordinate frame origin offset from the geocenter can also be estimated using the same set of data. Experience in this area has indicated that an over-constrained network, where more baselines or sites than necessary are fixed, can in fact produce a degraded solution. This is because in an over-constrained network the a priori uncertainty in the fixed parameters that are more than necessary will result in a suboptimal filter weighting. The solution will then be highly influenced by the mismodeling of these parameters.

In the first strategy proposed, the fiducial baselines are treated in three different ways:

(A) Fix two fiducial baselines.

(B) Constrain two fiducial baselines by a priori weighting.

(C) Fix only one fiducial baseline.

The baselines define the orientation of the adopted coordinate frame. The absolute scaling can be fixed either by the length of these baselines or by the Earth's gravitational constant, GM. Both are known to an accuracy of about one part in 10^8. The baseline length is used to define the absolute scaling so that the resulting coordinate frame will be consistent with the VLBI frame defined by the fiducial baselines. For the case with two baselines fixed, it is rather convenient to select one of the fiducial stations common to both fixed baselines as the reference site. The filter process is so designed that the baselines between the reference site and all other

nonfiducial sites are adjusted along with the Earth Orientation Parameters (EOP), namely polar motion (x_p, y_p) and UT1-UTC rate, the GPS satellite orbits, and the absolute coordinates of the reference site, which in turn infer the adjustment of the geocenter position coordinates. The Earth's GM is also adjusted, although the data strength may not be great enough to improve the value of GM appreciably.

In the second strategy, the same GPS tracking network of globally distributed stations is used. However, only the longitude of a reference site is held fixed; all other site coordinates are adjusted simultaneously with the GPS orbits. Here, the GM of Earth provides the absolute scaling. The station heights can be derived from the adjusted periods of GPS orbits and pseudorange measurements. The time signature of the measurements defines the latitude. Figure 2 graphically demonstrates the time signature of the measurements for two hypothetical cases. The first graph shows the periodic signature generated by the pseudorange (ρ) measurements to an orbiting GPS from a stationary receiver. The period is equal to the GPS orbit period, which is nearly 12 hours, and the amplitude is proportional to the geocentric position vector of the receiver projected onto the orbital plane. The second graph shows the case when a stationary GPS satellite is above the equator of a spinning Earth. The period is now 24 hours, and the amplitude is proportional to the cosine of the receiver latitude. The variation of the signature with respect to the receiver latitude is depicted in the sketch. Because of the difference in period, the effects due to the rotation of the receiver can be separated from the GPS orbiting signature and the latitude can unambiguously be solved.

A simple mathematical model can be written out for the estimate of geocenter offset. This offset is expressed as the weighted mean of the position offsets of all stations. The equations corresponding to the geocenter offset ΔG are represented as

$$\Delta G_x + \Delta x_i + v_i = 0, \qquad x \to y, z; \qquad i = 1, 2, \ldots, n \tag{8}$$

where Δx_i is the x component of the i^{th} geocentric station position offset and v_i is the error associated with Δx_i. The corresponding error covariance of the geocenter offset can be expressed as

$$\underset{3 \times 3}{\text{Var}\,(\Delta G)} = [\mathbf{A}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{A}]^{-1} \tag{9}$$

where

$$\underset{3 \times 3n}{\mathbf{A}^{\mathrm{T}}} = [-\mathrm{I} - \mathrm{I} \ldots - \mathrm{I}]$$

and W is a ($3n \times 3n$) weight matrix which is the inverted covariance matrix of the station position estimates.

IV. Covariance Analysis

A covariance analysis was carried out to assess the accuracy with which the geocenter offset from the origin of the adopted coordinate frame can be determined with each of the approaches proposed in previous sections. A full constellation of 18 GPS satellites distributed in six orbital planes was assumed. A data arc spanning over 34 hours from a network of six globally distributed tracking stations was also assumed. The three fiducial sites are the three NASA Deep Space Network (DSN) tracking sites (Fig. 3) at Goldstone, California; Canberra, Australia; and Madrid, Spain. The remaining sites in Japan, Brazil, and South Africa are nonfiducial sites. Simultaneous GPS P-code pseudorange and carrier phase measurements are made at all of these stations. The relative positions of the three DSN sites have been measured repeatedly by VLBI over many years and are known to an accuracy of about 3 cm. Goldstone was selected to be the reference site because of its common VLBI visibility with the other two DSN sites at Canberra and Madrid. P-code pseudorange and carrier phase data noise were assumed to be 5 cm and 0.5 cm, respectively, when integrated over 30 minutes and corrected for ionospheric effects by dual-frequency combination.

Carrier phase biases were adjusted with a large a priori uncertainty. Table 1 lists the error sources assumed for the first strategy. The robustness of the GPS measurements allows all the GPS and station clocks to be treated as white-noise processes and adjusted [3], [4] to remove their effects on the solutions. Also adjusted are the zenith tropospheric delays at all ground sites, which were treated as random-walk parameters to model the temporal change. Such models have been proved to be effective in removing their errors without seriously depleting the data strength [9].

The GPS covariance and simulation analysis software system, OASIS [15], recently developed at JPL, was used to carry out the study. In OASIS, partial derivatives with respect to cartesian components of site locations and the geocenter are readily produced. It is shown in the Appendix that baseline partials are related to site location partials as follows.

(1) The partial derivative with respect to a cartesian component of the reference site is the sum of all partial derivatives with respect to the same component of all sites forming the baselines. Note that this is also the partial derivative with respect to the same component of the geocenter position.

(2) The partial derivative with respect to a baseline cartesian component is the same as the partial derivative with respect to the same component of the nonreference site forming the baseline.

Hence, the site location coordinate partials can readily be used in place of the baseline coordinate partials, and the geocenter offset coordinate partials in place of the reference site absolute coordinate partials.

The second strategy assumes the same network of six tracking sites. The estimated quantities are the coordinates of all six sites except the longitude of the reference site (Goldstone), together with the GPS satellite states, white-noise clocks, random-walk troposphere parameters, and carrier phase biases. Because the longitude of Goldstone is held fixed, the position components need to be given in a geodetic coordinate system, viz., longitude, latitude, and height. Table 2 lists the assumption variations that apply to this strategy. Other assumptions are kept the same as in Table 1. With this strategy, the error covariance matrix of geocenter offset is given by Eq. (9) in the previous section.

V. Results of Covariance Analysis

In the covariance analyses for both strategies, data arcs of various lengths were used to study the solution convergence. In all cases the station at Goldstone was considered to be the reference site, although in the second strategy any of the ground sites can be a reference site where the only fixed component is the longitude.

Table 3 indicates the a priori error associated with the fiducial baselines, Goldstone-Canberra and Goldstone-Madrid, in all three cases of Strategy 1. The value of GM was adjusted, although it was found that the data strength of the GPS measurements is not great enough to improve on its a priori value. It should be noted that adjusting Earth's GM makes GPS satellite states consistent with the absolute scaling as implied by the baselines.

Figure 4 shows the total error of the origin offset as the length of the data span increases from 6 hours to 34 hours for Case A. At the end of 34 hours the origin offset error is 4.0 cm (rms of all three components). The bar chart shows a rapid reduction of error in origin offset between 6 and 12 hours. The result continues to improve after 12 hours but not at a very high rate. The reason for this can be seen in Fig. 5. After 12 hours the origin offset error has come down to the level of baseline error; data gathered thereafter only gradually reduces the effects of data noise. At the end of 34 hours the effect of data noise is reduced to 3.4 cm and would continue to reduce as the arc length increases. The contribution of the baseline error, however, dropped to about 2.5 cm after 12 hours and remained virtually unchanged thereafter. This indicates that the geocenter can be determined only up to the a priori accuracy of the fiducial baselines. Therefore, with this strategy, any improvement on the baseline accuracy can improve the accuracy of the origin offset from the geocenter. For instance, it is customary to find baselines reported with a higher accuracy in length than in the other two components. When a smaller error of 1 cm is assumed for the fiducial baseline length, along with 3 cm for the transverse and vertical components, the rms error on the origin offset from the geocenter reduces to 3.5 cm with a 34-hour arc of GPS measurements. Figure 6 shows the origin offset error for Case B, where the baseline vectors constrained to their a priori error are also estimated. The geocenter offset error after 34 hours reduces to 3.8 cm. Note that the error involved here is mainly due to data noise alone. Results from Case C, where the Goldstone-Madrid baseline is the only baseline fixed, are plotted in Figs. 7 and 8. The geocenter offset error after 34 hours is 4.4 cm, which is slightly worse than the previous cases. In Fig. 8, however, the effect due to the fixed baseline reduces to 2 cm after 12 hours and settles at 1.7 cm after 18 hours. The effect due to the data noise will continue to decrease for longer data arc, but the baseline effect will remain unchanged, as shown by Figs. 5 and 8. When the EOP are not estimated, the geocenter offset error after 34 hours is found to be 4.1 cm. This slight improvement is due to reduced data noise effect when fewer parameters are estimated.

In the second strategy no tracking site coordinates, except the longitude of the site at Goldstone, were held fixed. Here, as before, simultaneous adjustment of all GPS satellite states, tracking site coordinates, carrier phase biases, and zenith tropospheric corrections were carried out for various arc lengths ranging between 6 and 34 hours. Figure 9 plots the variation of the rms error of the origin offset from the geocenter with respect to the data arc length. The errors affecting the origin offset from the geocenter in this strategy are the data noise and the GM of Earth, which defines the absolute scaling. At the end of 6 hours the rms error of the origin offset is 143.7 cm, which reduces to 8 cm at the end of 12 hours. This indicates that the control on the absolute scaling and the orientation in latitude is greatly improved after all the GPS satellites have been tracked by the globally distributed sites for a complete orbit cycle. At the end of 34 hours the rms error reduces to 2.1 cm. The graph shows a strong trend of decreasing rms error as the arc length increases. This indicates that the origin offset determination is limited only by the data noise. This result can be compared with Case C of Strategy 1 when EOP are not estimated; there is about a 50% improvement in the geocenter offset error with this method. The Earth's GM is known accurately enough so that its effect is on the order of 0.2 cm after 12 hours and is 0.1 cm at the end of 34 hours.

In the analysis of Strategy 2, the effects of polar motion and UT1–UTC have not been included. However, GPS measurements are insensitive to any constant UT1–UTC bias error. The analyses done with different cases of Strategy 1 have indicated that a constant bias for polar motion and a UT1–UTC rate can be included in the filter as additional adjusted parameters without significantly degrading the performance.

VI. Effect of Coordinate Frame Origin Offset on Orbit Determination of Low Earth-Orbiting Satellite

To gain further insight into the significance of accurate definition of geocenter, the effect on the radial position of a low Earth-orbiting satellite, in particular TOPEX, was studied. The error assumptions used are the same as given in Table 1 except for those parameters listed in Table 4. The result presented by Case C of Strategy 1 shows that the origin offset accuracy is 4.4 cm (Fig. 7) with only one baseline fixed and a data arc of 34 hours. This value is the most pessimistic of all the results presented. Here, the origin offset was assumed to have an error of 4 cm in each component and left unadjusted. A reduced dynamic tracking technique [16] was implemented in the study where a fictitious 3-D force on TOPEX was adjusted as process noise with constrained a priori uncertainty. Figure 10 plots the error in the radial component of TOPEX caused by various sources. The total error in TOPEX altitude over the 2-hour arc has an rms value of 9.7 cm. Figure 11 shows the altitude error variation with time, along with the part contributed by a 4-cm geocenter uncertainty, over the 2-hour arc. Without the refinement with GPS measurements, the geocenter position uncertainty would be greater than 10 cm, and the TOPEX altitude determination error would be greater than 14 cm.

VII. Summary and Conclusions

A geocentric coordinate frame provides a practical global reference system with a physically meaningful and unambiguous definition of the coordinate origin. Two basic strategies for establishing a geocentric coordinate frame for GPS measurements have been investigated. All three cases of the first strategy make use of the precise relative positions which have been predetermined by VLBI to fix the frame orientation and the absolute scaling, while the offset from the geocenter is determined from GPS measurements. The reference frame thus adopted is consistent with the VLBI coordinate system. The second strategy establishes a reference frame by holding only the longitude of one of the tracking sites fixed. The absolute scaling is inferred by the adopted gravitational constant (GM) of Earth; the orientation in latitude is inferred from the time signature of Earth rotation in the GPS measurements. The coordinate system thus defined is a geocentric Earth-fixed coordinate system. The covariance analysis has shown that geocentric positioning to an accuracy of a few centimeters can be achieved with just a one-day arc of precise GPS pseudorange and carrier phase data.

Each of the two strategies has its advantages in different applications. The first strategy should be adopted in applications requiring a coordinate frame consistent with the VLBI reference frame. Among these applications are the monitoring of crustal motions in areas which have been investigated by VLBI observations and the determination of the Earth rotation parameters, viz., polar motion and variation of UT1–UTC. The second strategy, which holds the longitude at a reference site fixed, strictly limits itself in an ECEF frame established by the adopted values for the fixed longitude and the GM of Earth, and by GPS measurements. This method provides a superior result as long as the precise applications are within the same ECEF frame. Applications in which such an ECEF coordinate frame can be adopted include datum definition and network densification in an area where ECEF coordinates are appropriate. Various topographic and oceanographic surveys and prospecting surveys can benefit from its simplicity. In TOPEX orbit determination, this method can also be very convenient if a CTS frame such as WGS 84 is adopted.

References

[1] B. W. Parkinson and S. W. Gilbert, "NAVSTAR: Global Positioning System–Ten Years Later," *Proc. IEEE*, vol 71, no. 10, pp. 1177–1186, Oct. 1983.

[2] R. J. Milliken and C. J. Zoller, "Principles of Operation of NAVSTAR and System Characteristics," *Navigation*, vol. 2, no. 2, pp. 95–106, Summer 1978.

[3] T. P. Yunck, W. G. Melbourne, and C. L. Thornton, "GPS-Based Satellite Tracking System for Precise Positioning," *IEEE Trans. Geosci. & Remote Sensing*, vol. GE-23, no. 4, Jul. 1985.

[4] T. P. Yunck, S. C. Wu, and J. T. Wu, "Strategies for Sub-Decimeter Satellite Tracking with GPS," *Proc. 1986 IEEE Position Location and Navigation Symp.*, Las Vegas, NV, Nov. 1986.

[5] *Department of Defense World Geodetic System 1984*, DMA Tech. Rep. 8350.2, The Defense Mapping Agency, Sep. 1987.

[6] E. R. Swift, "NSWC's GPS Orbit/Clock Determination System," *Proc. First Int. Symp. on Precise Positioning with GPS*, Rockville, MD, pp. 51–62, Apr. 1985.

[7] J. M. Davidson, et al., "The March 1985 Demonstration of the Fiducial Concept for GPS Geodesy: A Preliminary Report," *Proc. First Int. Symp. on Precise Positioning with GPS*, Rockville, MD, pp. 603–611, Apr. 1985.

[8] O. J. Sovers, et al., "Radio Interferometric Determination of Intercontinental Baselines and Earth Orientation Utilizing Deep Space Network Antennas: 1971 to 1980," *J. Geophys. Res.*, vol. 89, no. B9, pp. 7597–7607, Sep. 1984.

[9] S. M. Lichten and J. S. Border, "Strategies for High-Precision Global Positioning System Orbit Determination," *J. Geophys. Res.*, vol. 92, no. B12, pp. 12751–12762, Nov. 1987.

[10] G. H. Born, R. H. Stewart, and C. A. Yamarone, "TOPEX–A Spaceborne Ocean Observing System," in *Monitoring Earth's Ocean, Land, and Atmosphere from Space–Sensors, Systems, and Applications*, A. Schnapf (ed.), AIAA, Inc., New York, NY, pp. 464–479, 1985.

[11] J. M. Dow and L. G. Agrotis, "Earth Rotation, Station Coordinates and Orbit Solutions from Lageos during the MERIT Campaign," *Proc. Int. Conf. on Earth Rotation and Terrestrial Reference Frame*, Columbus, OH, pp. 217–235, Jul. 1985.

[12] G. H. Kaplan, "The IAU Resolutions of Astronomical Constants, Time Scales, and the Fundamental Reference Frame," *USNO Circular*, no. 163, U. S. Naval Observatory, Washington, DC, Dec. 1981.

[13] I. I. Mueller, *Spherical and Practical Astronomy as Applied to Geodesy*, F. Ungar Publishing Co., Inc., 1969.

[14] W. G. Melbourne, et al., *MERIT Standards*, IAU/IUGG Joint Working Group on Rotation of Earth, Project MERIT, 1983.

[15] S. C. Wu and C. L. Thornton, "OASIS–A New GPS Covariance and Simulation Analysis Software System," *Proc. First Int. Symp. on Precise Positioning with GPS*, Rockville, MD, pp. 337–346, Apr. 1985.

[16] S. C. Wu, T. P. Yunck, and C. L. Thornton, "Reduced-Dynamic Technique for Precise Orbit Determination of Low Earth Satellites," AAS paper 87-410, AAS/AIAA Astrodynamics Specialists Conf., Kalispell, MT, Aug. 1987.

Table 1. Error sources and other assumptions for Strategy 1 (fixing baselines)

Reference site:	Goldstone
Other fiducial sites:	Canberra, Madrid
Nonfiducial sites:	Brazil, Japan, South Africa
GPS constellation:	18 satellites in 6 orbital planes
Cutoff elevation:	10 degrees
Data type:	P-code pseudorange; carrier phase
Data span:	6–34 hours
Data interval:	30 minutes
Data noise:	5 cm–pseudorange; 0.5 cm–carrier phase
Carrier phase bias:	10 km (adjusted)
Clock bias:	3 μsec–white noise (adjusted)
GPS epoch state:	10 m; 1 mm/sec (adjusted)
Geocenter position:	10 m each component (adjusted)
Baseline coordinates:	3 cm each component–fiducial; 10 cm each component–nonfiducial (adjusted)
Zenith troposphere:	Random walk parameter (adjusted): 20 cm bias; 1.3 cm batch to batch
Earth's GM:	One part in 10^8
Solar pressure:	10%
(UT1–UTC) rate:	10 m/day (adjusted)
Polar motion (x_p, y_p):	10 m (adjusted)

Table 2. Variations of assumptions from Table 1 for Strategy 2 (fixing only one longitude)

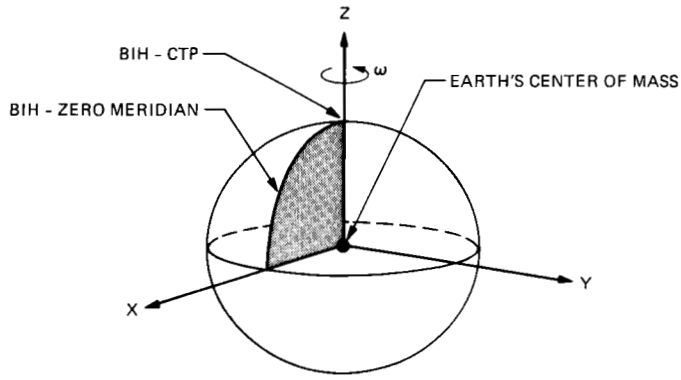
Reference site:	Goldstone
Reference site coordinates: (adjusted)	10 m (latitude) 0 m (longitude) 10 m (height)
Other site coordinates: (adjusted)	10 m each component

Table 3. Fiducial baselines in Strategy 1

Case	Baselines	Adjusted	a priori
A	Goldstone-Canberra	no	3 cm
	Goldstone-Madrid	no	3 cm
B	Goldstone-Canberra	yes	3 cm
	Goldstone-Madrid	yes	3 cm
C	Goldstone-Canberra	yes	10 cm
	Goldstone-Madrid	no	3 cm

Table 4. Variations of assumptions from Table 1 for TOPEX orbit determination

Data span:	2 hours
Data interval:	5 minutes
TOPEX epoch state:	1 km; 1 m/sec (adjusted)
3-D force on TOPEX:	Process-noise (adjusted): 0.50 $\mu m/s^2$ bias; 0.35 $\mu m/s^2$ batch to batch
Gravity:	50% of current uncertainty (20 X 20 lumped)
Geocenter:	4 cm each component



ORIGIN: EARTH'S CENTER OF MASS

Z AXIS: PARALLEL TO DIRECTION OF THE CONVENTIONAL TERRESTRIAL POLE (CTP) FOR POLAR MOTION AS DEFINED BY THE BIH ON THE BASIS OF THE COORDINATES ADOPTED FOR THE BIH STATIONS

X AXIS: INTERSECTION OF THE ZERO MERIDIAN PLANE DEFINED BY BIH (WGS 84 REFERENCE MERIDIAN PLANE) AND THE PLANE OF CTP EQUATOR

Y AXIS: MEASURED IN THE PLANE OF CTP EQUATOR, $\pi/2$ EAST OF X AXIS, THUS COMPLETING THE RIGHT-HANDED EARTH-CENTERED, EARTH-FIXED (ECEF) ORTHOGONAL COORDINATE SYSTEM

Fig. 1. The WGS 84 coordinate system.

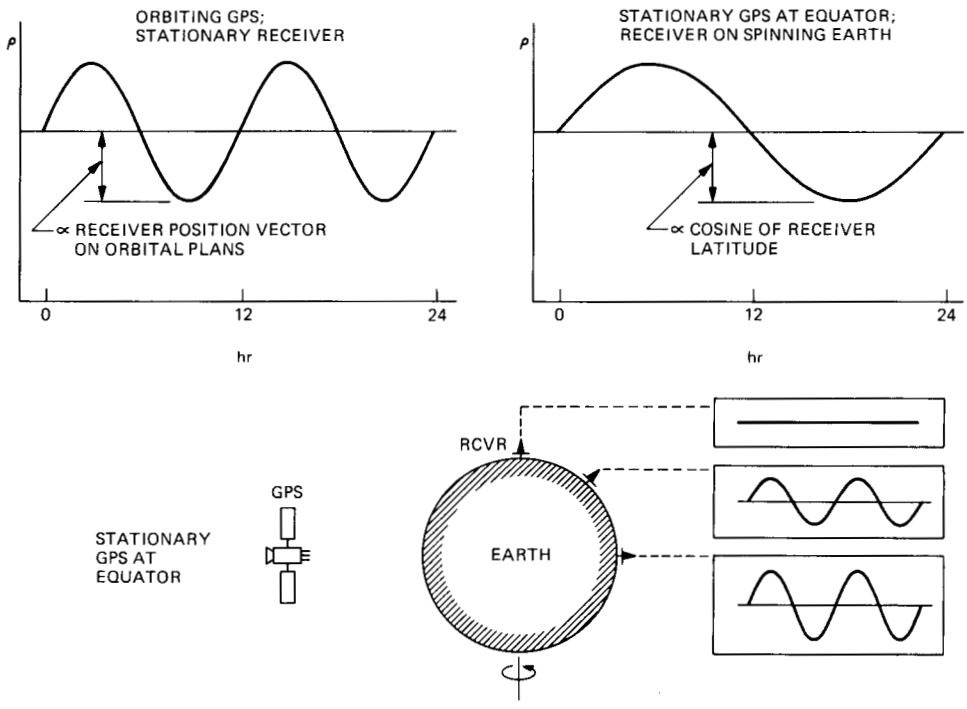


Fig. 2. Time signature of GPS measurements.

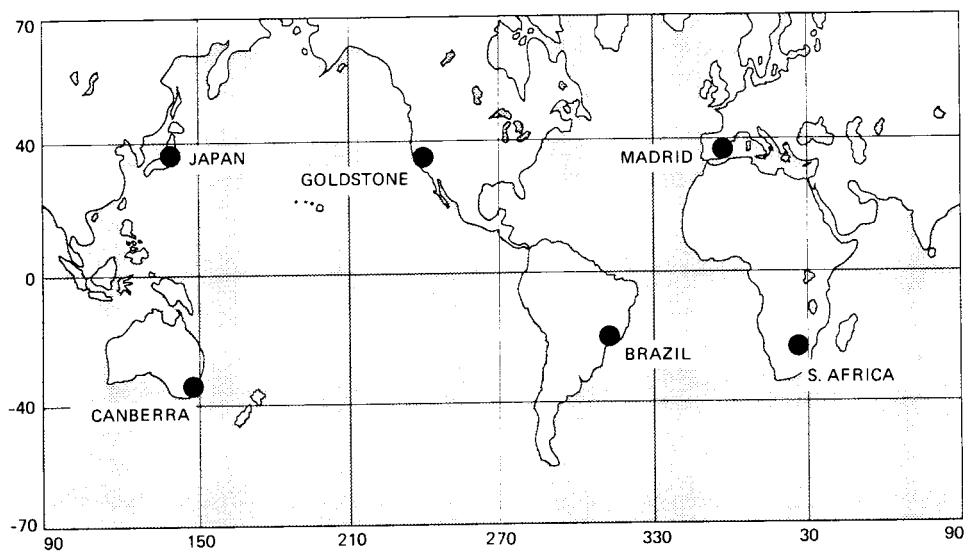


Fig. 3. A global GPS tracking network.

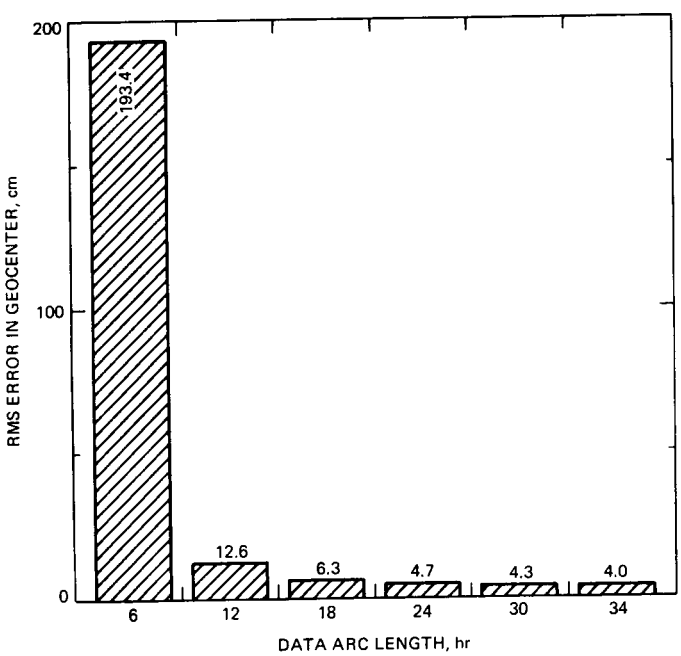


Fig. 4. Convergence of geocenter offset determination using Case A of Strategy 1 (two baselines fixed).

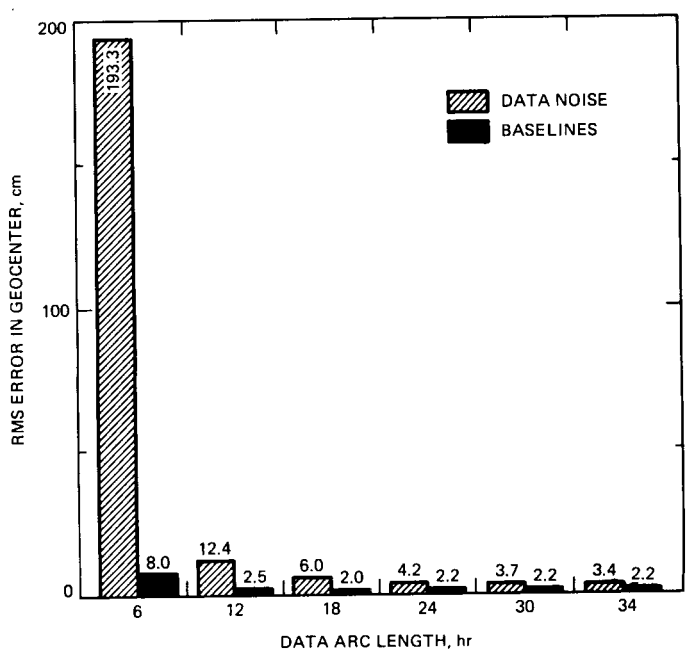


Fig. 5. Breakdown of geocenter offset determination error using Case A of Strategy 1 (two baselines fixed).

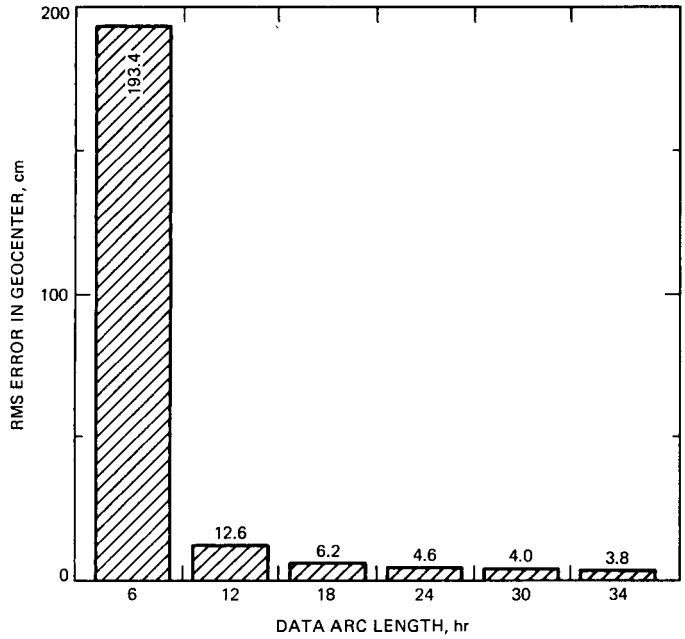


Fig. 6. Convergence of geocenter offset determination using Case B of Strategy 1 (two constrained baselines).

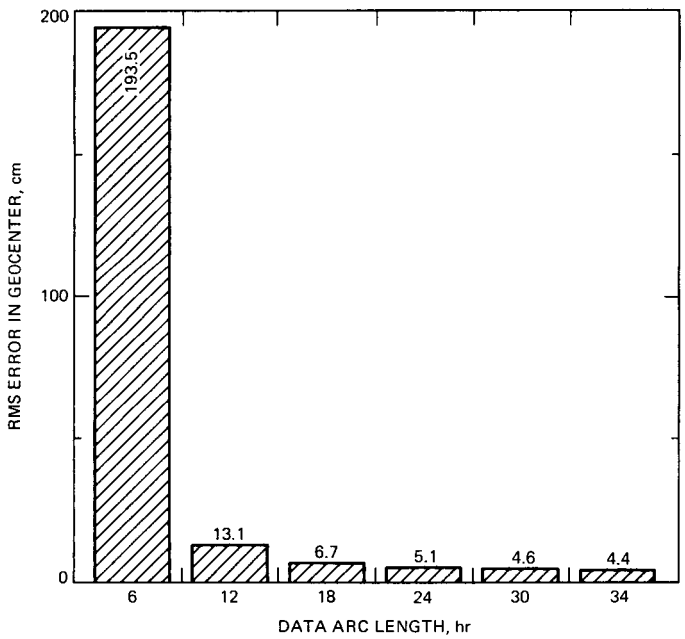


Fig. 7. Convergence of geocenter offset determination using Case C of Strategy 1 (one baseline fixed).

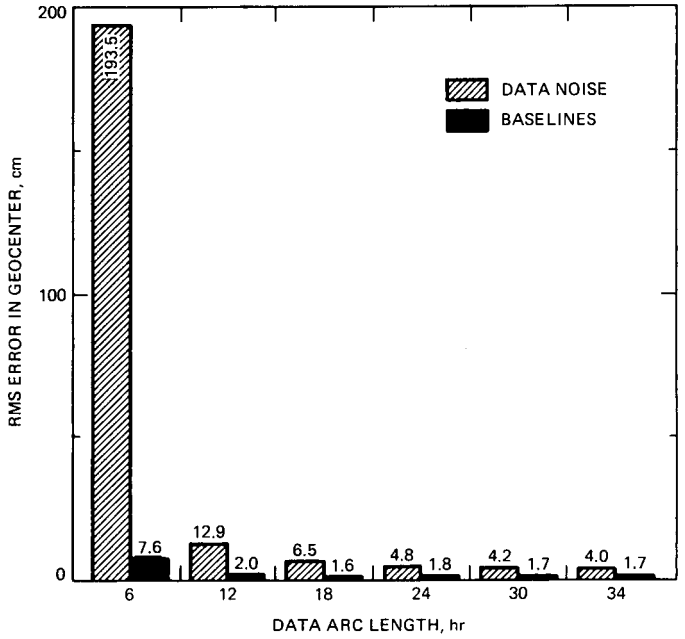


Fig. 8. Breakdown of geocenter offset determination error for Case C of Strategy 1 (one baseline fixed).

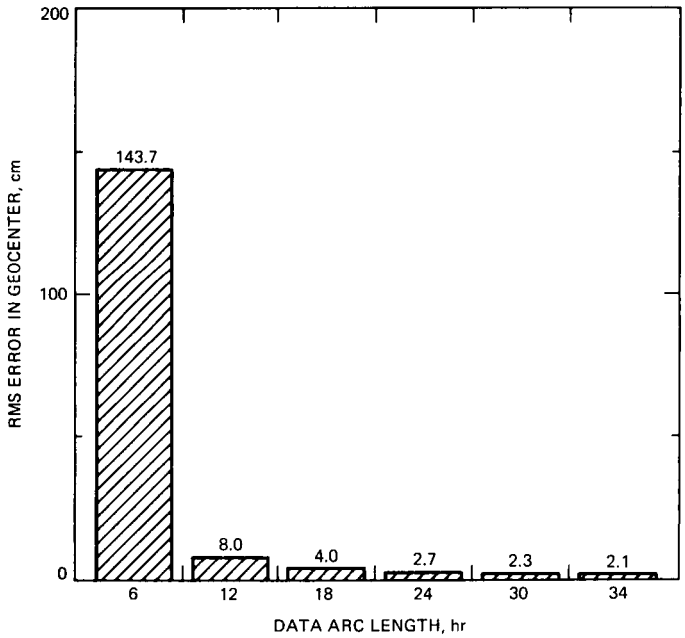


Fig. 9. Convergence of geocenter offset determination using Strategy 2 (longitude at Goldstone fixed).

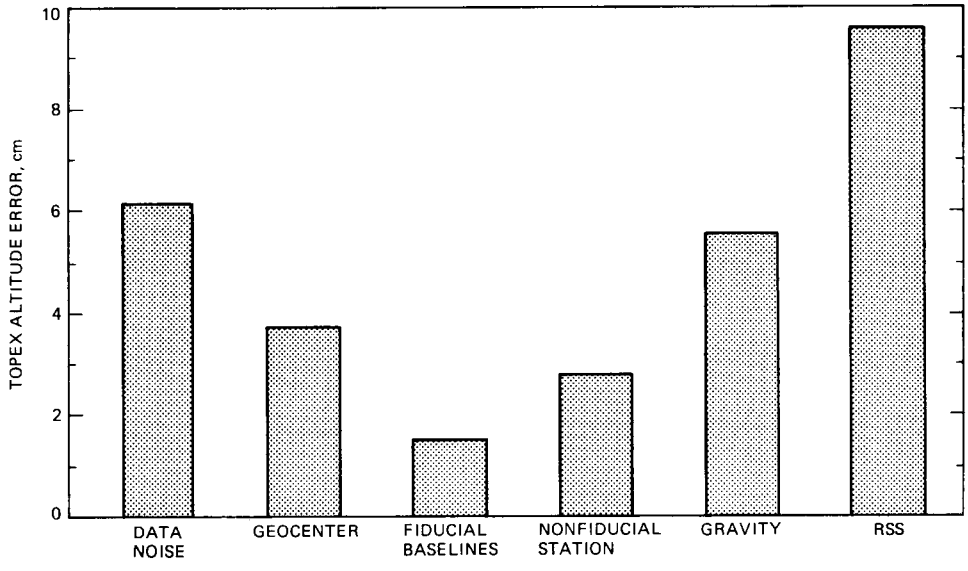


Fig. 10. Breakdown of TOPEX altitude determination error.

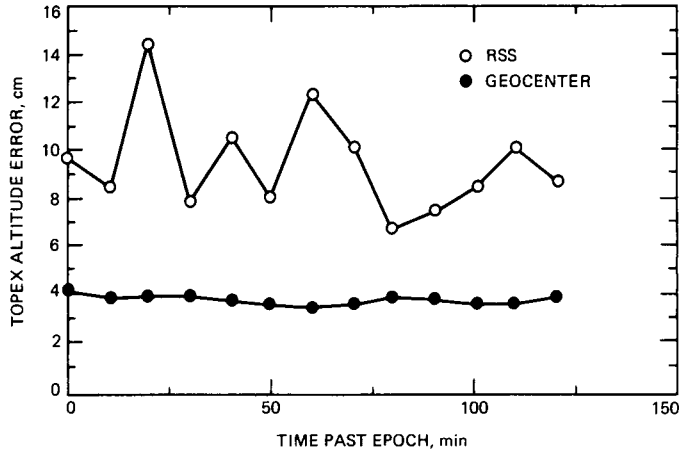


Fig. 11. Total TOPEX altitude error and effects of 4-cm geocenter error over a 2-hour period.

Appendix

Measurement Partial Derivatives with Respect to Baseline Components

Let the cartesian coordinates of the set of N tracking sites be $(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_N, y_N, z_N)$. We can form the following baseline components:

$$b_{x,j} = x_j - x_1, \qquad x \to y, z; \quad j = 2, 3, \ldots, N \tag{A-1}$$

where site 1 has been selected as the reference site with which all baselines are formed. For completeness, we also define

$$b_{x,1} = x_1, \qquad x \to y, z \tag{A-2}$$

for the reference site. For simplicity, but without loss of generality, partial derivatives with respect to only the x-component of baselines will be derived. The relation for the other two components follows directly. These equations can be rearranged as

$$x_j = \begin{cases} b_j, & j = 1 \\ b_j + b_1, & j = 2, 3, \ldots, N \end{cases} \tag{A-3}$$

from which the following partial derivative can be formed:

$$\frac{\partial x_i}{\partial b_{x,j}} = \begin{cases} 1, & j = 1 \\ \delta_{ij}, & j = 2, 3, \ldots, N \end{cases} \tag{A-4}$$

where δ_{ij} is the Kronecker delta. The partial derivative of a measurement R with respect to the baseline components b_j can be expressed in terms of those with respect to the site coordinates x_j by the following chain rule:

$$\frac{\partial R}{\partial b_{x,j}} = \frac{\partial R}{\partial x_1}\frac{\partial x_1}{\partial b_{x,j}} + \frac{\partial R}{\partial x_2}\frac{\partial x_2}{\partial b_{x,j}} + \ldots + \frac{\partial R}{\partial x_N}\frac{\partial x_N}{\partial b_{x,j}} \tag{A-5}$$

which, with the substitution of Eq. (A-4), becomes

$$\frac{\partial R}{\partial b_{x,j}} = \begin{cases} \sum_{n=1}^{N} \frac{\partial R}{\partial x_n}, & j = 1 \\ \frac{\partial R}{\partial x_j}, & j = 2, 3, \ldots, N \end{cases} \tag{A-6}$$

Hence, the partial derivative of the measurement with respect to a cartesian component of a baseline is the same as that with respect to the same component of the nonreference site forming the baseline; and the partial derivative with respect to a component of the reference site is the sum of all partial derivatives with respect to the same component of all sites forming the baselines.

Determination of GPS Orbits to Submeter Accuracy

W. I. Bertiger and S. M. Lichten
Tracking Systems and Applications Section

E. C. Katsigris
Geology and Planetology Section

Orbits for satellites of the Global Positioning System (GPS) have been determined with submeter accuracy. Tests used to assess orbit accuracy include orbit comparisons from independent data sets, orbit prediction, ground baseline determination, and formal errors. One satellite tracked for 8 hours each day shows rms errors below 1 m even when predicted more than 3 days outside of a 1-week data arc. Differential tracking of the GPS satellites in high Earth orbit provides a powerful relative positioning capability, even when a relatively small continental U. S. fiducial tracking network is used with less than one-third of the full GPS constellation. To demonstrate this capability, baselines of up to 2000 km in North America were also determined with the GPS orbits. The 2000-km baselines show rms daily repeatability of 0.3 to 2 parts in 10^8 and agree with very-long-baseline interferometry (VLBI) solutions at the level of 1.5 parts in 10^8. This GPS demonstration provides an opportunity to test different techniques for high-accuracy orbit determination for high Earth orbiters. The best GPS orbit strategies included data arcs of at least 1 week, process noise models for tropospheric fluctuations, estimation of GPS solar pressure coefficients, and combined processing of GPS carrier phase and pseudorange data. For data arcs of 2 weeks, constrained process noise models for GPS dynamic parameters significantly improved the solutions.

I. Introduction

The Global Positioning System (GPS), expected to be fully operational by the early 1990s, will consist of 24 satellites evenly spaced in six orbit planes at an altitude of about 20,000 km. Knowledge of GPS orbits will provide the basis for highly accurate ground and satellite user positioning. A wide variety of users will benefit from this positioning capability. GPS will be used at NASA's Deep Space Network (DSN) stations in conjunction with very-long-baseline interferometry (VLBI) radiotelescopes for atmospheric calibrations, precise ground station position determination, monitoring of Earth orientation changes on time scales of less than one day [1], and possibly time synchronization at the nanosecond level. Differential GPS-based accuracies for high Earth orbiters are expected at the several-meter level for altitudes of 5,000 to 40,000 km [2]. Spacecraft maneuvering near and docking with the Space Station will carry GPS receivers and will use GPS signals for real-time and near-real-time navigation and guidance. Low Earth orbiting spacecraft such as TOPEX/Poseidon [3] and the Earth Observing System platforms will have orbit determination available in post-real time to an

accuracy of better than 10 cm with kinematic smoothing techniques using advanced receivers to track GPS satellites simultaneously with a worldwide GPS ground tracking network. The relatively low cost and convenience of GPS ground receivers have created many new opportunities to monitor cm-level crustal motions in geologically active regions. Very dense ground networks may achieve accuracies equaling or surpassing those available from other generally more restrictive techniques such as VLBI or satellite laser ranging.

The GPS applications with the most stringent requirements include the DSN applications, subdecimeter low Earth orbit determination, cm-level measurements of Earth crustal motion, and cm-level monitoring of changes in Earth orientation. To reach these goals, GPS orbits will have to be determined to better than 50-cm accuracy. The Jet Propulsion Laboratory has been developing and testing GPS orbit estimation software and techniques for several years with the goal of demonstrating the capability for high-accuracy orbit determination. GPS data from several field experiments in 1985 and 1986 have been used to determine precise GPS orbits. Although only seven developmental GPS satellites were operational and the ground fiducial tracking network was limited to sites in the continental United States, covariance studies indicated that with currently available GPS receivers and antennas it should be possible to produce orbits for well-tracked GPS satellites accurate to 1 m. Achieving this 1-m accuracy capability is a major milestone on the road to ultrahigh-precision GPS applications.

In this article we present results demonstrating submeter accuracy for the GPS orbits determined from these field tests. Ground station coordinates were estimated simultaneously along with the orbit parameters. Accuracy of better than 3 cm has been achieved over baselines up to 2000 km, proving that GPS is already a very powerful technique for precise positioning over continental distances.

II. Data Acquisition and Processing

A series of GPS field experiments took place in March and November 1985, June 1986, and January 1988. These experiments were organized by JPL and were cooperative ventures with several different organizations participating. In the March 1985 experiment, data were collected for about 1 week at ground sites in the continental United States only. In November 1985, the tracking network also included three sites in Mexico, and the experiment lasted for about 2 weeks. The June 1986 experiment covered a 3-week period and included sites in the Caribbean region as well as a dense network of stations in Southern California. The data from the experiments up through 1986 has been processed, with the January 1988 data expected to be distributed shortly. This article reports analysis based on the data collected during the November 1985 and June 1986 experiments. Results from the March 1985 experiment have been reported earlier [4], [5].

A. The November 1985 and June 1986 Data Sets

The November 1985 GPS experiment took place from November 12 through November 24. The June 1986 experiment spanned about 3 weeks; the results presented in this article are based on the first part of the experiment from June 2 to June 10. Figure 1 shows the locations of the ground tracking sites, which represent a subset of the total number that participated in the experiments. TI 4100 GPS receivers [6] were operated at most of the sites. SERIES-X(7) receivers[1] built at JPL were used at Mojave and at Owens Valley Radio Observatory (OVRO). In November 1985, water vapor radiometers (WVRs) were available for wet tropospheric delay calibrations at OVRO, at the Mexican sites, and at Mojave for part of the experiment. In June 1986, WVRs were used at Haystack, Mojave, and some of the Caribbean sites. Dry tropospheric delay calibrations were computed from surface measurements of barometric pressure. For receivers located at sites where WVRs were not available, wet tropospheric delay corrections were computed from surface meteorology data. Single-day-arc baseline and GPS orbit solutions were generated for examination of residuals and to check the quality of the data. November 12 and 17 were excluded because of data outages and other difficulties at some of the fiducial sites. For multiday-arc solutions, the November 1985 data set, covering 12 days, was divided first into two arcs of 7 and 5 days (November 13–19 and November 20–24). This was necessary due to a maneuver of more than 100 km which took place on November 20 during which Navstar 4 was moved to a new orbit. Eventually, we modeled and solved for the maneuver as described below, and a single long data arc covering November 13–24 was constructed. During these GPS experiments, periods of common ground visibility lasted about 6 to 8 hours. Most satellites were visible continuously from a given ground station for only about 3 hours, so several times during the tracking period the receivers switched to a new combination. Navstar 8 was unusual in that it was visible for up to 8 hours from most of the ground sites.

Because of the short tracking periods (relative to the GPS orbital period of 12 hours) from a limited network of ground sites, orbits determined from single-day passes in the 1985 and 1986 field tests were significantly weaker than those determined from multiday arcs. The additional strength from the multiday arcs derives mainly from the visibility of the satel-

[1] R. B. Crow, F. R. Bletzacker, R. J. Najarian, G. H. Purcell, J. I. Statman, and J. B. Thomas, "Series-X Final Engineering Report," JPL Internal Document D-1476, Jet Propulsion Laboratory, Pasadena, California, 1984.

lites from the ground during multiple orbit revolutions. With observations over more than one revolution, the orbital periods are more accurately determined and therefore the positions of the orbital nodes are more precise. The down-track satellite components benefit especially from the multiday arcs, as is sometimes manifested by a corresponding improvement in the eastern component baseline accuracy [5]. A second advantage of multiday-arc solutions is the $\sqrt{n}$ improvement in precision, where n is the number of measurements, which can apply to any orbital or baseline component. As more satellites are launched and the tracking network expands geographically, shorter arcs will achieve the same level of orbit accuracy.

B. Data Collection and Processing

Both GPS carrier phase and pseudorange data were received at all the sites equipped with TI receivers. Carrier phase only was used from the Series-X receivers, which are codeless. The carrier signals at L1 and L2 bands (1.227 and 1.575 GHz) are modulated by a pseudorandom noise code called the P code, which operates at 10.23 MHz. Continuously tracked GPS carrier phase provides a very precise measure of *range change*, while the P code provides a measure of *absolute range*. The pseudorange is considerably noisier than the carrier phase data type, and in this experiment the pseudorange was corrupted by errors due to multipath signals. The GPS observables at the two L-band frequencies are linearly combined to remove the portion of ionospheric delay which varies as the inverse square of the frequency. For more details about the characteristics of the GPS signals, see [7] and [8]. Hereafter, the terms "carrier phase" and "pseudorange" refer to these linear combinations of L1, L2 and P1, P2. The GPS data were processed with the GPS Inferred Positioning SYstem (GIPSY) orbit determination and baseline estimation software, which was completed and tested at JPL shortly after the data were collected for the 1985 experiments.

III. Orbit Determination Approach

The JPL orbit determination software utilizes a pseudoepoch-state U-D factorized Kalman filter[2] [9]. The filter works as a batch sequential program with the option to model parameters as first-order exponentially correlated process noise, also commonly called colored noise. The GIPSY software uses the J2000 reference system with observation partials for parameters computed relative to the satellite epoch states [10].

For the November 1985 data set, the nominal GPS ephemerides were obtained by using broadcast orbits as initial values and iterating to improve those orbits (and remove large undesired offsets resulting from different coordinate frame conventions) with a small subset of the data. During the iteration, GPS solar radiation pressure coefficients were also determined and these coefficients and the new orbits were used as nominal values for the more comprehensive precise orbit filter runs. In June 1986, the hybrid postfit ephemeris from the Naval Surface Weapons Center (NSWC) was used for nominal trajectories and a similar iteration was performed prior to the final filter solutions.

The high-precision orbit determination strategy used with the November 1985 and June 1986 GPS data was based on previous experience with the March 1985 GPS experiment [5]. A key aspect of the orbit estimation process is the fiducial concept, where three or four receivers with well-known coordinates in a consistent reference frame are held fixed while all other parameters, including orbital states and coordinates of the nonfiducial sites, are estimated simultaneously in the filter. Reference 5 and the references therein discuss the fiducial concept as well as alternative approaches. The fiducial receivers for the experiments discussed in this article were collocated at VLBI sites. Haystack, Richmond, and Fort Davis were generally treated as fiducials. In some runs OVRO was held fixed, and in others Hat Creek was used as a fiducial so that one of the normally fixed receivers could be estimated using the GPS data in order to test the internal consistency. As more GPS satellites are launched and the ground tracking network is expanded from North America to include stations on other continents, we expect that fewer fiducial constraints will need to be applied and more station location parameters will be determined from the GPS data.

A. Clock and Bias Parameters

The results presented here are based on estimation of station and GPS clocks as white process noise. At each measurement epoch, each active clock is assumed to have a value uncorrelated to its value at other epochs. Although some station clocks were running off hydrogen masers and most of the GPS clocks have well-characterized behavior typical of rubidium and cesium atomic standards, this extra information was not used. The white noise clocks were estimated simultaneously with the other parameters in the filter. When the process noise model for clocks is white noise, the results are virtually the same as would be achieved with double differencing [11] for clock elimination.

The GPS carrier phase, when continuously tracked, provides a very precise measure of range change from measurement epoch to measurement epoch. The absolute phase, however, and hence the absolute range from transmitter to receiver, is ambiguous by an integral number of wavelengths. For each station-satellite tracking pass, a carrier phase bias parameter is

[2]S. C. Wu, W. I. Bertiger, J. S. Border, S. M. Lichten, R. F. Sunseri, B. G. Williams, P. J. Wolff, and J. T. Wu, "OASIS Mathematical Description, v. 1.0," JPL Internal Document D-3139, Jet Propulsion Laboratory, Pasadena, California, 1986.

estimated, along with the other adjusted parameters, ignoring the integer constraint on its value. Over a period of hours, the signature of the *range change* precisely measured with the carrier phase enables the orbit to be determined. The pseudorange, on the other hand, provides a more direct range determination but is much noisier and more susceptible to multipath errors than the carrier phase data type. The ultimate accuracy would be reached if *carrier range* were available. Carrier range is a range determined from carrier phase with the bias ambiguities all resolved; it has the best features of both carrier phase (low noise, low multipath) and pseudorange (absolute range measure and geometric strength). Successful carrier phase ambiguity resolution has been reported over baselines of up to 2000 km with *bias fixing* or *bias optimizing* techniques applied to single-day arcs [12], [13]. The results reported here were achieved by estimation of the white noise clock and carrier phase bias parameters with a very large a priori uncertainty, so that their solutions were basically unconstrained. Bias fixing was not used to resolve the carrier phase ambiguities.

B. Tropospheric Delay Fluctuations

The troposphere was modeled as a spherical shell which adds a delay along the GPS signal path:

$$\rho = \rho_{z_d} R_d(\theta) + \rho_{z_w} R_w(\theta) \tag{1}$$

where ρ_z is the zenith tropospheric path delay and R is an analytic mapping function [14] to map zenith delays to line-of-sight path delays at elevation θ. There are two components to the tropospheric delay—the wet and the dry, denoted here with subscripts w and d. The dry component can be determined under the assumption of hydrostatic equilibrium using the ideal gas law for dry air to better than 1 cm [15]. The wet delay component, although considerably smaller than the dry, exhibits greater time and spatial variation and is much more difficult to determine accurately.

WVRs were operated at some of the GPS tracking sites alongside the GPS receivers. WVR calibrations are believed accurate to 2 cm or better for determination of the wet zenith path delay [15]. The algorithm described in [16] was used to determine these calibrations. At the other sites, surface meteorology (SM) measurements (temperature, air pressure, and relative humidity) were used for the wet zenith delay calibration [17]. The SM calibration is much less reliable than the WVR calibration because the surface meteorological conditions are not always well correlated with the total atmospheric water vapor content [15].

Wet zenith delay corrections to the calibrations (WVR or SM) were estimated with the GPS data for all GPS tracking sites. For most sites with WVRs, a constant wet zenith delay parameter with an a priori constraint of 3 cm was estimated for each 8-hour tracking day. For sites using SM calibrations, the wet zenith delay was estimated daily with a 20-cm a priori constraint. In addition, stochastic residual delays were estimated for SM sites in order to remove signatures which could result from temporal variations in the troposphere, time-varying errors in the SM calibrations, errors in the mapping function, or spatial inhomogeneities due to azimuthal asymmetry in the water vapor content. In some cases, tightly constrained process noise troposphere residual delay parameters were also estimated for the WVR sites.

C. Troposphere Process Noise Models

The stochastic model in the GIPSY filter is for a first-order exponentially correlated process noise [9]. The measurements are processed in discrete time segments, known as *batches*. In each batch, process noise parameters are modeled as piecewise constants. At the end of a batch, a process noise time update adds noise to the covariance matrix and thus causes the time-varying behavior of the stochastic parameters. The process noise time update for the jth batch maps the estimates and covariance for the stochastic parameters into batch $j+1$:

$$\mathbf{p}_{j+1} = \mathbf{M}_j \mathbf{p}_j + \mathbf{w}_j \tag{2}$$

where $\mathbf{p}_j$ is a vector of estimates for the stochastic parameters and $\mathbf{M}$ is a diagonal process noise mapping matrix. The process noise $\mathbf{w}_j$ is a random process with zero mean and

$$E\left(\mathbf{w}_j \mathbf{w}_k^T\right) = \mathbf{Q}\delta_{jk} \tag{3}$$

where $\mathbf{Q}$ is the covariance matrix diagonal and δ_{jk} is the Kronecker delta function [9]. The diagonal entries of $\mathbf{M}$ are given by

$$m_{ij} = \exp\left[-(t_{j+1} - t_j)/\tau_{ij}\right] \tag{4}$$

where t_j is the start time for the jth batch and τ_{ij} is the time constant for the ith stochastic parameter at the jth batch. The corresponding diagonal entry in the matrix $\mathbf{Q}$ is

$$q_{ij} = \left(1 - m_{ij}^2\right) \sigma_{iss}^2 \tag{5}$$

where σ_{iss}, the steady-state sigma for the ith stochastic parameter, is the noise level that would be reached if the system were left undisturbed for a time much greater than τ. The process noise model for each parameter is fully specified by σ_{ss} and τ, which can also vary with time, although the subscript j has been left off σ_{ss} for simplicity of notation. There are two special limiting cases: white process noise, and a ran-

dom walk. For white process noise, $\tau = 0$, $m = 0$, and, as can be seen in Eq. (2), the a priori covariance for the process noise parameters, **p**, is completely reset at the end of each batch, including zeroing of off-diagonal terms and inserting q for the variance on the diagonal. For white noise, the process at each time step is independent and uncorrelated with the process at other time steps. The opposite case is the random walk. Here both σ_{ss} and τ are unbounded, since a steady state is never reached and $\tau = \infty$. For the random walk, however, q is still defined by Eq. (2), where **M** is now equal to the identity matrix. The Allan variance [18], $\sigma_A^2(\Delta t)$, which is often used to characterize clock and atmospheric fluctuations [19], is directly related to the random walk q:

$$\sigma_A^2(\Delta t) = q/\Delta t^2 \quad \text{(random walk)} \tag{6}$$

A wide range of process noise models has been tested on the March and November 1985 and June 1986 GPS data sets [5], [15], [20]. The random walk zenith tropospheric delay models with $\sqrt{q}$ in the range 2 to 4×10^{-7} km·s$^{-1/2}$ (for SM sites) produced the best daily baseline repeatability, agreement with VLBI, and orbit repeatability. When only constant zenith delay parameters were estimated, orbit and baseline accuracies were worse by about a factor of 2. The value of $\sqrt{q}$ adopted for most of the November 1985 and June 1986 analyses was 2×10^{-7} km·s$^{-1/2}$, corresponding to about 6 cm variation over a 24-hr period. There is evidence from VLBI residuals [19], [21] that tropospheric delay and delay-rate fluctuations can be well modeled as random walks for Δt greater than a few hundred seconds. Since the GPS data were compressed to 300-sec intervals, the use of random walk tropospheric fluctuation models for GPS is consistent with the VLBI findings. As discussed in [5], however, the GPS process noise parameters were used to estimate a fluctuating *residual* correction to calibrated data (WVR or SM), and it is not clear that this quantity will have the same stochastic characteristics as the tropospheric fluctuations themselves. It was assumed that the spectral characteristics of both the tropospheric fluctuations and the residuals after calibration would be similar.

D. Solar Radiation Pressure and Other Nongravitational Forces

Multiday arcs covering 1 to 2 weeks were used to achieve the highest GPS orbit accuracy. With a global tracking system equipped with high-performance GPS receivers and a full 24-satellite GPS constellation, covariance analysis predicts GPS orbit accuracy well below 1 m after just several hours of tracking [22]. However, as can be seen from Fig. 1, the ground tracking network during 1985 and 1986 was limited, with fiducials in North America only. The seven GPS satellites that were operating at that time by design tend to converge over the southwest United States. This further limited common ground visibility to less than 8 hours per day and reduced the geometrical strength of the system. In addition, the pseudorange available from the TI receivers that were used in 1985 and 1986 was highly contaminated by ground multipath, thereby raising the effective measurement noise. Because of these factors, multiday arcs were necessary to achieve the desired improvement in ephemeris accuracy.

With single-day (8-hour) orbit solutions, there is very little sensitivity to errors in the GPS solar radiation pressure coefficients [4]. However, for multiday arcs with multiple orbit revolutions, the orbital period is much more accurately determined and therefore the results become sensitive to downtrack errors resulting from integration of accelerations due to solar radiation forces acting on the GPS satellites, mismodeled solar radiation, or unmodeled forces such as thermal radiation from the spacecraft body. The GPS Block I ROCK4 model was used to represent accelerations resulting from solar radiation pressure. As described in [5] and references therein, ROCK4 models 13 surfaces on the satellites according to their size, curvature, reflectivity, specularity, and absorption characteristics. The model as implemented in GIPSY allows for adjustment of three parameters: G_x, G_y, and G_z. G_x and G_z are scaling factors in the local spacecraft x and z directions, where the z axis is positive along the antenna toward the center of the Earth, the y axis is along the solar panel support beam normal to the spacecraft-Sun-Earth plane, and the x axis completes a right-handed coordinate system. G_y represents a constant acceleration in the y-axis direction, often referred to as the y-bias parameter [23].

In principle, the G_x and G_z parameters should have the same value if the spacecraft were perfectly aligned and the model were correct. GPS orbits have been determined in the past with estimation of only G_y and one parameter (designated here as G_{xz}) to represent both G_x and G_z [24], [25]. For the multiday arcs determined with the March 1985 data [5], G_x, G_y, and G_z were estimated independently as constants for each satellite with the intention of adding an extra degree of freedom ($G_x \neq G_z$) to absorb unmodeled accelerations and known deficiencies in the ROCK4 model, which are thought to amount to as much as 4 m error over a 14-day prediction interval [23]. This strategy was successful for data arcs up to about 1 week long. However, for longer arcs of up to 2 weeks, a noticeable and systematic degradation in daily baseline repeatability occurred with the three-parameter constant solar coefficient approach. A new approach was adopted in which two constant solar pressure coefficients were estimated, G_y and G_{xz}, along with two tightly constrained process noise parameters, G_x and G_z. With this approach, daily baseline repeatability continued to improve as the arc was lengthened.

As an alternative to the use of process noise on G_x and G_z, small colored noise accelerations were introduced for the long November 13–24 arcs. These consisted of three parameters representing a constant thrust in the directions of down-track, cross-track, and altitude.

Both stochastic solar pressure parameters and stochastic thrust parameters gave similar repeatability results for the 2-week data arc, Fig. 2. The results were insensitive to the time constant, τ, as long as the value of q was held about constant (see Eqs. 4 and 5). For G_x and G_z, τ was varied from 1 to 28 days with $q \approx 9.3 \times 10^{-6}$, and the repeatability results were essentially the same. The results for the thrust parameters were more sensitive to the time constant. The best results were not obtained until the time constant was made less than 2 days and the value of $q \approx 2 \times 10^{-28}$ km^2/sec^4. Similar results were obtained when the time constant for the thrust parameters was made as short as 0.5 day. A single stochastic thrust in the direction of the satellite long track also yielded good repeatability. With the present data, which consists of relatively short arcs at the same time each day, one cannot determine whether the effects are due to solar pressure mismodeling or other random or systematic accelerations acting on the GPS spacecraft. One possibility is that a solar pressure stochastic model is physically the correct approach but the stochastic fictitious thrust model works as well, as long as the thrust time constant is short enough to absorb the daily variations implicit in the solar radiation pressure signature. Note that another approach to this problem has been proposed in which fictitious force parameters with 24-hr resonances are estimated to remove solar pressure, gravity, and other dynamic errors which tend to repeat daily [28]. It is hoped that the sources of these forces can be isolated when more data and global tracking are available. This may be possible with the CASA UNO data set [29].

E. GPS 4 Maneuver

A maneuver was performed on GPS 4 (SV 8) at approximately UT 0320 November 20, 1985. In a few hours the spacecraft state was changed by over 100 km and it was moved to a slightly different orbit. This time was in the middle of the experiment but outside the data collection interval, which was from about UT 1100–1900 each day. In order to perform continuous orbit determination over the entire experiment interval (November 13–24) for all the satellites simultaneously, we had to model and estimate the maneuver. The NSWC parameterized the maneuver as a constant thrust over 5 minutes. We used a four-parameter model with three instantaneous velocity changes and a time-of-burn parameter, which we refer to as an impulsive motor burn. The impulsive motor burn allows for an instantaneous change in position and velocity while leaving the acceleration unchanged. Let $\Delta\mathbf{V}$ be the vector velocity increment, with three components ΔV_H, ΔV_C, and ΔV_L, where subscripts H, C, and L refer to the local spacecraft coordinates of altitude, cross-track, and down-track. Let ΔT_B denote the time of burn, e.g., the interval over which the burn is applied. Then the vector change in position is

$$\Delta\mathbf{r} = \frac{1}{2}\Delta\mathbf{V}T_B = \frac{1}{2}\mathbf{a}\Delta T_B^2 \tag{8}$$

The acceleration, $\mathbf{a}$, defined by $\Delta\mathbf{V}$ and ΔT_B, is the equivalent constant acceleration that would be experienced over the interval ΔT_B. The solution for the maneuver parameters ΔV_H, ΔV_C, ΔV_L, and ΔT_B was obtained by iterating over a data arc spanning November 18–22. The maneuver solution converged after three iterations and showed meter-level agreement with a separate solution for the GPS 4 position using only data collected after the maneuver (November 20–24). This nominal representation for the maneuver was further refined in later longer-arc high-precision GPS orbit solutions. It is interesting to note that the maneuver solution did not converge initially unless we constrained $G_x = G_z$ for GPS 4.

IV. Assessment of Orbit Accuracy

Five criteria were used to assess the accuracy of the GPS orbits determined from multiday arcs in November 1985 and June 1986: (1) orbit repeatability; (2) orbit prediction; (3) daily baseline repeatability; (4) agreement between GPS and VLBI-determined baselines; and (5) formal errors from the orbit filter.

A. Orbit Repeatability

Orbit repeatability indicates the precision and, to some extent, the accuracy of the GPS orbits. Figure 3 illustrates how orbit repeatability was computed for the November 1985 experiment. The purpose of orbit repeatability is to compare orbits determined independently without any common measurements and then compute the rms difference over a time interval during which no data were used for either of the two solutions. Figure 4 shows the mean of the rms computed for all seven GPS satellites over a 6-hr interval on November 17. Figure 4 also shows the significant improvement attained when pseudorange is processed with the carrier phase and stochastic tropospheric delay models are used. From the formal errors, it appears that the pseudorange contributed little geometric strength to the orbit solutions, since the rms scatter of individual measurements was 100 to 300 cm for 6-minute measurement intervals, due mostly to ground multipath. However, when the pseudorange and carrier phase are processed together and a common clock is estimated, the pseudorange provides a priori knowledge of the clock and carrier phase bias parameters at the several-nanosecond (100–300 cm) level, significantly improving the orbits.

GPS 6 and 8, the two satellites with the most data and best ground viewing geometry, had formal errors below 1 m for most of the November experiment and had significantly lower orbit repeatability rms than the other satellites. Figure 5 shows the repeatability computed over a 24-hr interval on November 17 when no measurements were used for either orbit solution. Since the two solutions being compared were determined from independent data sets (Fig. 3), it is concluded that submeter orbit precision has been demonstrated for these two satellites.

B. Orbit Prediction

Orbit prediction is a stringent test of orbit accuracy, since the estimated spacecraft position and velocity are mapped outside of the measurement interval to give a predicted satellite state. Orbit errors tend to be magnified in this mapping process. The accuracy of the orbit models used for mapping are also tested by the orbit prediction test, in addition to the accuracy of the satellite ephemerides. Figure 6 shows the prediction test we applied to our November 1985 multiday-arc orbits. The average rms errors for four of the satellites were 0.7 m, 0.8 m, and 1.7 m in altitude, cross-track, and downtrack components. GPS 8, which was tracked longer than any of the other satellites, had a prediction rms error well below 1 m, even when mapped more than 3 days into the second arc. For the first 6 hours of the prediction interval, the rms error was 50 cm or less for all three position components. These results are shown in Figs. 7 and 8.

C. Daily Baseline Repeatability and Agreement Between GPS and VLBI Solutions

To further assess the GPS orbit accuracy determined with multiday arcs, we have examined daily baseline repeatability as well as agreement with independent VLBI baseline measurements over continental distances (1000 to 2000 km). For the November 1985 experiment, we examined baselines between Hat Creek, CA and Fort Davis, TX (1933 km); Mojave, CA and Fort Davis, TX (1314 km); and Richmond, FL and Haystack, MA (2046 km). Daily repeatability was computed as the rms about the weighted mean of the daily baseline solutions determined simultaneously with one multiday-arc orbit solution. For these baselines in North America with good common visibility of the GPS, the rms scatter was 0.3 to 2 parts in 10^8 of baseline length for all vector components. Agreement with VLBI was 0.3 to 1.5 parts in 10^8 for baselines with the same type of GPS antenna at both ends. Figures 9 and 10 show the results for the 2000-km baselines.

For determination of the Hat Creek–Fort Davis baseline, Fort Davis, Richmond, and Haystack were held fixed as fiducial stations. For the Richmond–Haystack baseline solution, a separate filter run was made with Hat Creek, Fort Davis, and Haystack fixed as the fiducial reference points. Note that in the case of the Richmond–Haystack baseline the Hat Creek fiducial does not have data in the second half of the 2-week data arc, reducing the number of fiducials to two for the second half. The data from the first half has already determined the GPS positions sufficiently that the lack of a third fiducial does not degrade the quality of the solution. Although high-quality orbits are a prerequisite for good precision and accuracy over long baseline distances, there are other factors that can affect baseline accuracy aside from GPS orbits. For example, local site vectors between the GPS and VLBI antennas sometimes are inaccurate by several cm. One such local survey error was recently discovered at OVRO, leading to a 5-cm discrepancy between the GPS and VLBI Mojave–OVRO baseline until it was corrected [26]. Therefore, while the daily baseline repeatability provides a measure of consistency for orbits determined from multiday arcs, agreement with VLBI is a measure of *overall system accuracy*, which depends on a number of factors in addition to orbit accuracy.

When different GPS antennas were used at the ends of a baseline, although daily repeatability was excellent, agreement with VLBI was worse by 1 to 7 cm, with no apparent dependence on baseline length. This was noticed in both the November 1985 and June 1986 experiments. However, baselines with the same type of GPS antennas showed good agreement with VLBI. Since ephemeris errors tend to scale with baseline length, it was hypothesized that these discrepancies were due to local phenomena rather than orbital effects. Attention has been directed at phase center variations in the antennas, since the TI antennas are designed so that in operation the phase center variations nearly cancel out between sites that are not more than a few thousand km apart [6]. However, the Series-X antennas do not have the same phase center characteristics as the TI antennas, and the signatures resulting from the several-cm phase center variations that have been measured [27] could corrupt baseline measurements between unlike antennas. Therefore we qualify our high-precision results with the warning that measurements between different GPS antennas may be much less reliable and may be affected by unpredictable effects.

D. Baseline Repeatability Outside the Fiducial Network

The baselines in North America are fairly well determined because they are either inside or near the fiducial network and because the current limited GPS constellation by design is optimized for North America, especially the southwest United States. The formal errors from the filter are consistent with the results in Figs. 7, 8, and 9, predicting precision of 1 to 4 cm over these 2000-km baselines. To further test the robustness of the multiday-arc GPS orbits, we have determined baselines between Richmond and several sites in the Caribbean Sea

region occupied during the June 1986 experiment. Figure 10 shows daily baseline repeatability for Richmond–Grand Turk (1049 km) and Richmond–Isabela (1582 km) determined from an 8-day orbit fit for June 2–10, 1986. Because the Caribbean sites are far to the southeast of both the fiducial network and the optimal region for the GPS constellation, the formal errors for these baselines are typically 2 to 7 cm, somewhat worse than those of the North American baselines. Despite the degraded geometry and reduced common visibility, the baselines to the Caribbean sites show precision of 1 to 4 parts in 10^8. Some of the Caribbean sites were equipped with WVRs; for these as well as the sites without WVRs, residual tropospheric corrections determined from the GPS data were critical in achieving these levels of baseline precision. Since the Caribbean sites are considerably more humid than most of the North American sites, various strategies for reducing systematic errors due to uncertainties in the wet troposphere correction are being studied with this data set.

V. Conclusions

It has been demonstrated that submeter GPS orbits can be determined using multiday arc solutions with the current GPS constellation subset visible for about 8 hours each day from North America. Submeter orbit accuracy was shown through orbit repeatability and orbit prediction. North American baselines of 1000 to 2000 km in length can be estimated simultaneously with the GPS orbits to an accuracy of better than 1.5 parts in 10^8 (3 cm over a 2000-cm distance) with a daily precision of 2 parts in 10^8 or better. The most reliable baseline solutions are obtained using the same type of receivers and antennas at each end of the baseline. Baselines longer than 1000 km between Florida and sites in the Caribbean region have also been determined with daily precision of 1 to 4 parts in 10^8. The Caribbean sites are located well outside the fiducial tracking network and the region of optimal GPS common visibility, so these results further demonstrate the robustness of the multiday-arc GPS orbit solutions.

Process noise models have been used in the orbit determination filter to minimize systematic errors which can seriously affect ephemeris and baseline accuracy. These systematic effects include tropospheric delay fluctuations and small, unmodeled spacecraft accelerations. The process noise troposphere models improved all orbit and baseline solutions, regardless of length of data arc. Tightly constrained process noise representation for part of the solar pressure model significantly improved baseline repeatability for arcs longer than 1 week; however, an equally effective technique had fictitious thrusts estimated stochastically for each GPS satellite. Because of the limited ground visibility with the current constellation, it is not yet possible to determine whether the accelerations are genuinely related to solar radiation pressure or are due to other random or systematic forces acting on the spacecraft.

This demonstration of several-cm accuracy over distances of a few thousand km, despite a limited ground tracking network and a constellation of only seven satellites, proves that GPS provides a very powerful relative positioning capability. It shows that GPS techniques have the intrinsic data strength and robustness needed for DSN high-precision applications, as well as low Earth orbiter tracking and crustal motion studies.

References

[1] A. P. Freedman and J. O. Dickey, "Usefulness of GPS for the Precise Determination of Earth Orientation Parameters," *EOS*, vol. 68, no. 44, p. 1245, November 3, 1987.

[2] S. C. Wu, "Differential GPS Approaches to Orbit Determination of High-Altitude Earth Satellites," AAS/AIAA Astrodynamics Specialist Conference, paper AAS 85-430, August 12–15, 1985, Vail, CO, published in *Astrodynamics 1985*, vol. 58 of *Advances in the Astronautical Sciences*, pp. 1203–1220, 1986.

[3] S. M. Lichten, S. C. Wu, J. T. Wu, and T. P. Yunck, "Precise Positioning Capabilities for TOPEX Using Differential GPS Techniques," AAS/AIAA Astrodynamics Specialist Conference, paper AAS 85-401, August 12–15, 1985, Vail, CO, published in *Astrodynamics 1985*, vol. 58 of *Advances in the Astronautical Sciences*, pp. 597–614, 1986.

[4] J. M. Davidson, et al., *The Spring 1985 High Precision Baseline Test of the JPL GPS-Based Geodetic System*, JPL Publication 87-35, November 15, 1987.

[5] S. M. Lichten and J. S. Border, "Strategies For High Precision GPS Orbit Determination," *J. Geophys. Res.*, vol. 92, pp. 12751–12762, November 10, 1987. November 10, 1987.

[6] D. J. Henson, E. A. Collier, and K. R. Schneider, "Geodetic Applications of the Texas Instruments TI 4100 GPS Navigator," *Proceedings First International Symposium on Precise Positioning with GPS-1985*, vol. I, pp. 191–200, National Geodetic Information Center, NOAA, Rockville, MD, 1985.

[7] R. J. Milliken and C. J. Zoller, "Principle of Operation of NAVSTAR and System Characteristics," *Navigation*, vol. 25, pp. 95–106, 1978.

[8] E. H. Martin, "GPS User Equipment Error Models," *Navigation*, vol. 25, pp. 201–210, 1978.

[9] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, Orlando, Florida: Academic Press, 1977.

[10] O. J. Sovers and J. S. Border, *Observation Model and Parameter Partials for the JPL Geodetic GPS Modeling Software*, GPSOMC, JPL Publication 87-21, September 15, 1987.

[11] Y. Bock, S. A. Gourevitch, C. C. Counselman III, R. W. King, and R. I. Abbot, "Interferometric Analysis of GPS Phase Observations," *Manuscr. Geod.*, vol. 11, pp. 282–288, 1986.

[12] D. Dong and Y. Bock, "GPS Network Analysis: Ambiguity Resolution," *EOS*, vol. 69, no. 16, p. 325, April 19, 1988.

[13] G. Blewitt, "Successful GPS Carrier Phase Resolution for Baselines up to 2000 km in Length," *EOS*, vol. 69, no. 16, p. 325, April 19, 1988.

[14] G. Lanyi, "Troposphere Calibration in Radio Interferometry," *Proceedings of the International Symposium on Space Techniques for Geodynamics*, pp. 184–195, IAG/COSPAR, Sopron, Hungary, July 1984.

[15] D. M. Tralli, T. H. Dixon, and S. A. Stephens, "The Effect of Wet Tropospheric Path Delays on Estimation of Geodetic Baselines in the Gulf of California Using the Global Positioning System," *J. Geophys. Res.*, vol. 93, pp. 6465–6557, June 10, 1988.

[16] S. E. Robinson, "A New Algorithm for Microwave Delay Estimation From Water Vapor Radiometer Data," *TDA Progress Report 42-87*, vol. July–September 1986, pp. 149–157, Jet Propulsion Laboratory, Pasadena, CA, November 15, 1986.

[17] C. C. Chao, "A New Method to Predict Wet Zenith Range Correction from Surface Measurements," *JPL Technical Report 32-1526, vol. XIV, The Deep Space Network*, pp. 33–41, Jet Propulsion Laboratory, Pasadena, CA, 1974.

[18] D. W. Allen, "Statistics of Atomic Frequency Standards," *Proc. IEEE*, vol. 54, pp. 221–230, 1966.

[19] A. R. Thompson, J. M. Moran, G. W. Swenson, *Interferometry and Synthesis in Radio Astronomy*, New York: John Wiley & Sons, 1986.

[20] E. C. Katsigris, D. M. Tralli, and T. H. Dixon, "Estimation of Wet Tropospheric Path Delays in GPS Baseline Solutions for the 1986 Caribbean Experiment," *EOS*, vol. 69, no. 16, p. 324, April 19, 1988.

[21] R. N. Treuhaft and G. E. Lanyi, "The Effect of the Dynamic Wet Troposphere on Radio Interferometric Measurements," *Radio Science*, vol. 22, pp. 251–265, 1987.

[22] W. G. Melbourne, G. Blewitt, S. M. Lichten, R. E. Neilan, S. C. Wu, and B. E. Schutz, "Establishing a Global GPS Tracking System for Fiducial Control and Ephemeris Production," *EOS*, vol. 69, p. 323, no. 16, April 19, 1988.

[23] H. F. Fliegel, W. A. Feess, W. C. Layton, and N. W. Rhodus, "The GPS Radiation Force Model," *Proceedings First International Symposium on Precise Positioning with GPS-1985*, vol. I, pp. 113–120, National Geodetic Information Center, NOAA, Rockville, MD, 1985.

[24] E. R. Swift, "NSWC's GPS Orbit/Clock Determination System," *Proceedings First International Symposium on Precise Positioning with GPS-1985*, vol. I, pp. 51–62, National Geodetic Information Center, NOAA, Rockville, MD, 1985.

[25] R. I. Abbot, Y. Bock, C. C. Counselman III, and R. W. King, "GPS Orbit Determination," *Proceedings Fourth International Geodetic Symposium on Satellite Positioning*, Defense Mapping Agency and National Geodetic Survey, vol. I, pp. 271–273, Austin, TX, April 1986.

[26] J. Ray, "MOTIES Results," presented at the Crustal Dynamics Investigators Meeting, Jet Propulsion Laboratory, Pasadena, CA, March 22, 1988.

[27] A. Kleusberg, "GPS Antenna Phase Centre Variations," *EOS*, vol. 67, no. 44, p. 911, Nov. 4, 1986.

[28] O. L. Colombo, "Precise Determination of GPS Orbits and Station Positions," presented at IUGG Symposium, Vancouver, BC, August 1987.

[29] R. E. Neilan, et al., "CASA UNO GPS–A Summary of the January '88 Campaign," *EOS*, vol. 69, no. 16, p. 323, April 19, 1988.

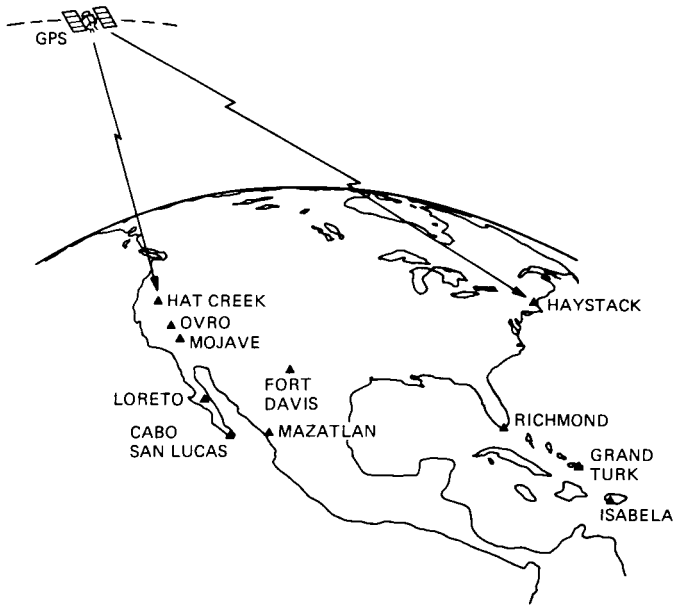


Fig. 1. Locations of ground tracking sites used in the analysis of GPS data from the November 1985 and June 1986 experiments.

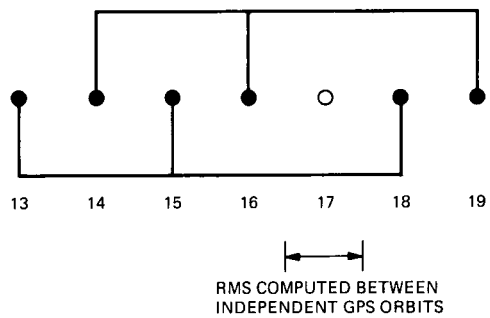


Fig. 3. Orbit repeatability for November 1985 uses data from Nov. 13, 15, 18 for one solution and Nov. 14, 16, 19 for the other. Rms difference between the two solutions is computed over Nov. 17, during which no data was taken.

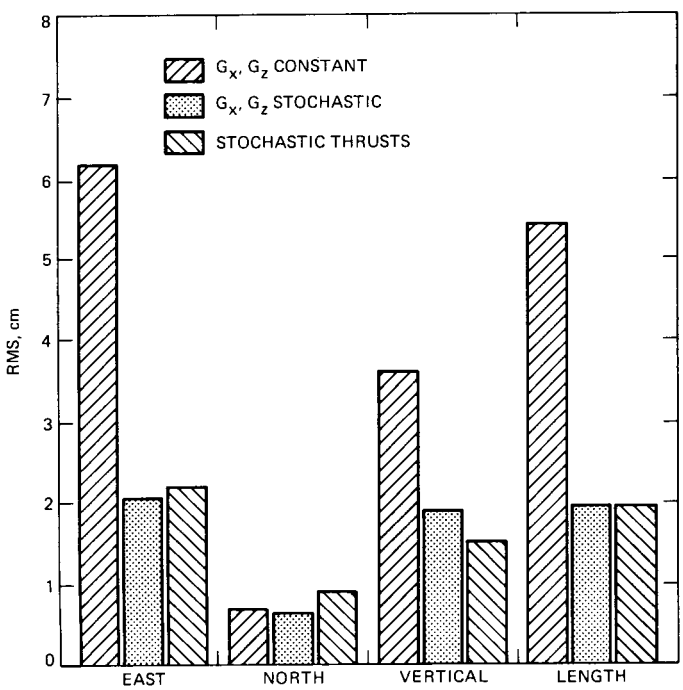


Fig. 2. Daily baseline repeatability with 2-week orbit arcs (Nov. 13-24, 1985) for the Mojave-Fort Davis 1314-km baseline showing dramatic improvement when stochastic force parameters are estimated for GPS satellites.

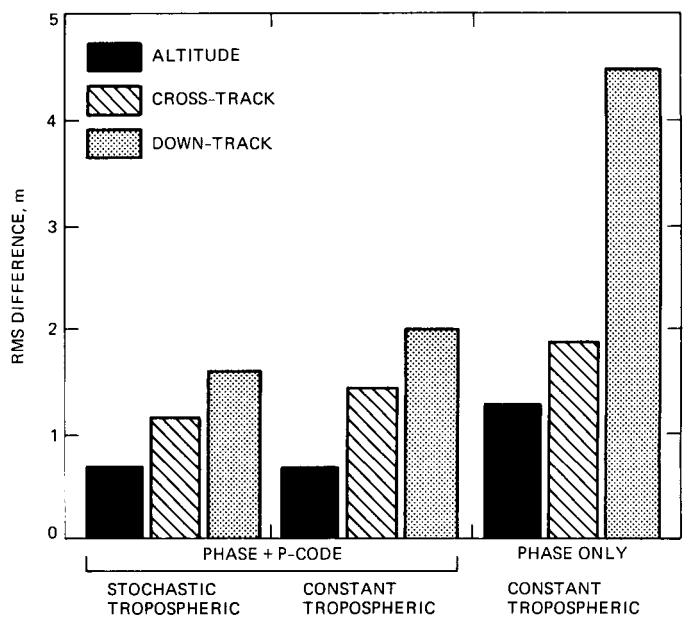


Fig. 4. Rms orbit repeatability over a 6-hr interval on Nov. 17. Rms between independent arcs shown in Fig. 3 for altitude, cross-track, and down-track components have been averaged for all seven GPS satellites.

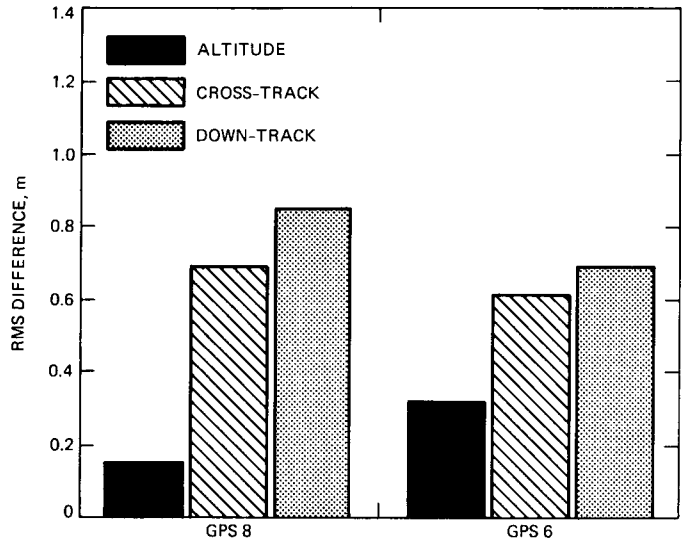


Fig. 5. Submeter orbit repeatability for GPS 6 and 8. Rms difference between two independent solutions is computed over a 24-hr interval on Nov. 17 when no measurements were used.

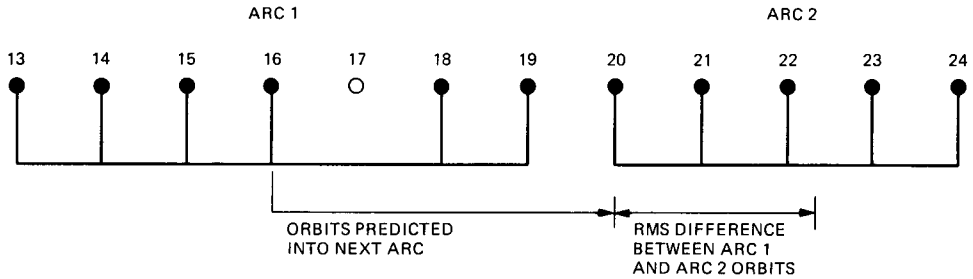


Fig. 6. Orbit prediction test for Nov. 1985. Orbits determined from Nov. 13-19 are mapped ahead, and the rms difference is computed with an independent solution determined from Nov. 20-24.

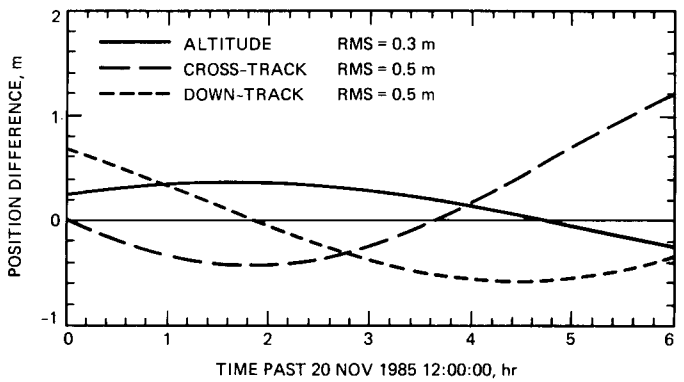


Fig. 7. Difference between arc 1 predicted orbit and arc 2 orbit as shown in Fig. 6 for GPS 8. Rms is taken over a 6-hr interval.

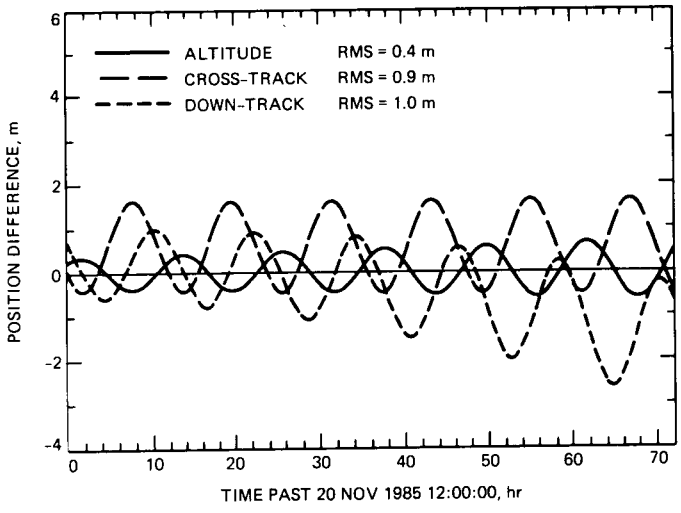


Fig. 8. Orbit prediction difference similar to Fig. 7 except that here the difference and rms are shown for a prediction interval of more than 3 days.

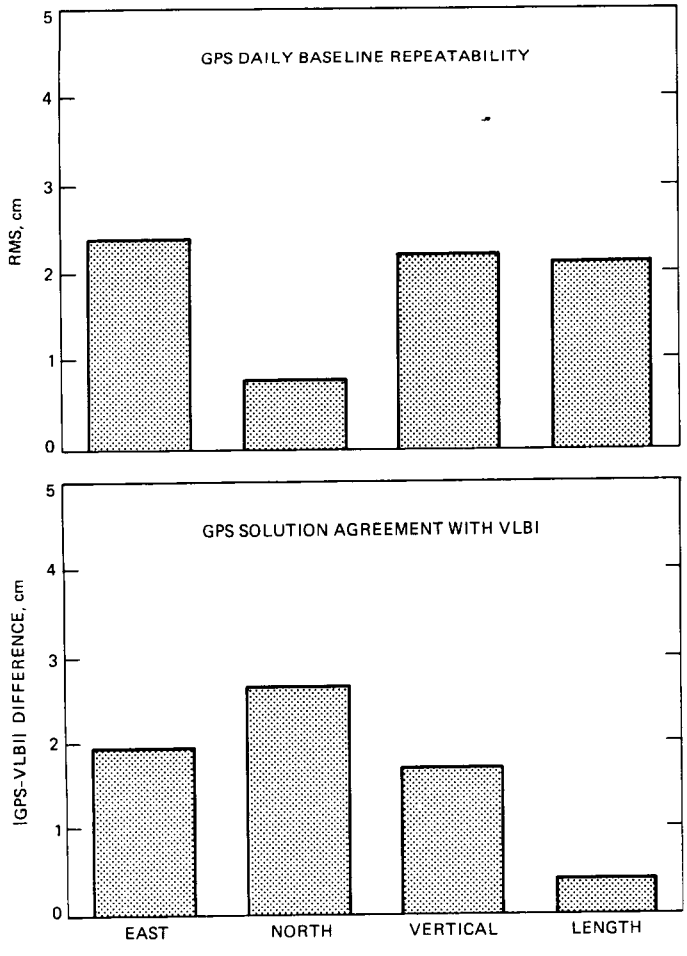


Fig. 9. Daily GPS baseline repeatability and agreement with VLBI for the Hat Creek-Fort Davis 1933-km baseline determined with a multiday orbit fit from Nov. 13-24, 1985.

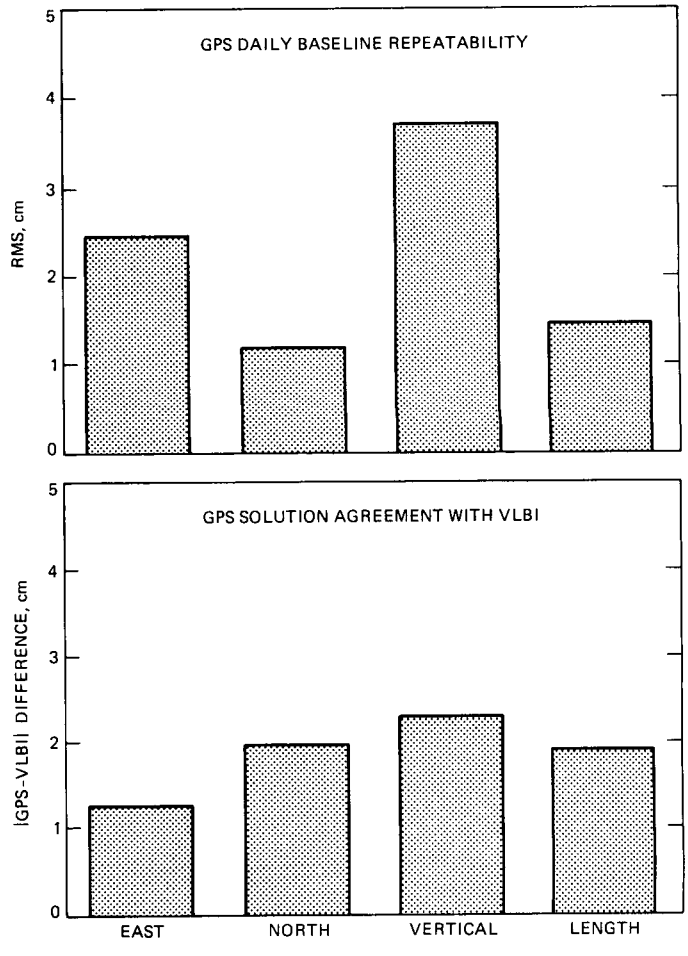


Fig. 10. Daily GPS baseline repeatability and agreement with VLBI for the Richmond-Haystack 2046-km baseline determined with a multiday orbit fit from Nov. 13-24, 1985.

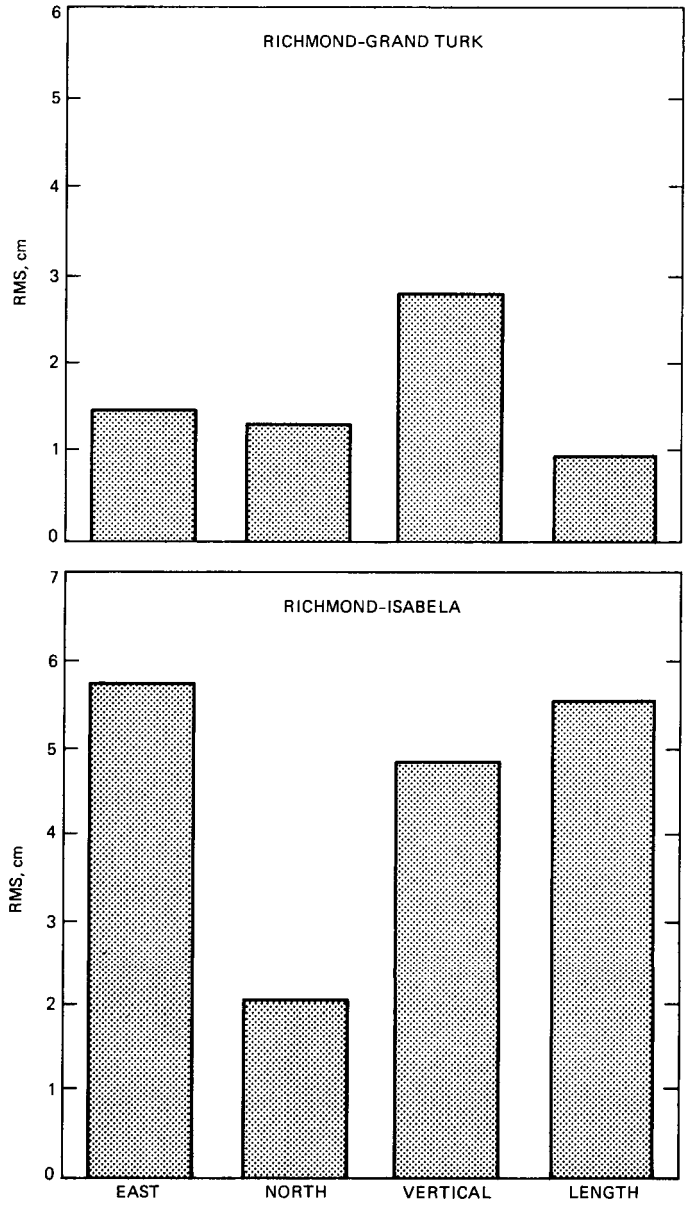


Fig. 11. Daily GPS baseline repeatability for two baselines (1049 and 1582 km) to the Caribbean region determined with multiday arc orbits from the June 1986 experiment.

Two-Way Coherent Doppler Error Due to Solar Corona

P. W. Kinman and S. W. Asmar
Telecommunications Systems Section

This report considers two-way coherent Doppler error resulting from phase scintillations induced on the uplink by the solar corona. It is shown that this error can be estimated by taking statistics on the differential Doppler measurements. Typical estimates for the error are given for four Sun-Earth-probe angles and for integration times ranging from 1 second to 1 minute. These results are based on data collected during the 1985 Voyager 2 solar conjunction.

I. Introduction

When a spacecraft is in the same area of the sky as the Sun and beyond it, the radio beams to and from that spacecraft experience phase scintillations due to passage through the solar corona. These scintillations will dominate all other sources of error for two-way coherent Doppler measurement when the Sun-Earth-probe angle is small. Thus, for those missions that feature numerous or long-lasting solar conjunctions, especially inner planetary missions, it is important to characterize this source of Doppler error. Unfortunately, the relevant parameters of the solar corona are highly variable, and adequate statistical characterizations do not exist. It has, however, been possible to measure the phase scintillations induced on 2.3- and 8.4-GHz downlinks by the solar corona. It is in fact possible to make this measurement using any Sun-encountering spacecraft with two downlinks of different but related frequencies. The phase scintillations induced on the downlinks, once measured, can be removed from the Doppler phase record. The phase scintillations induced on the uplink, on the other hand, remain. These cannot be measured or removed unless two simultaneous uplinks are available.

This report considers two-way coherent Doppler error resulting from phase scintillations induced on the uplink by the solar corona. This error is estimated by taking statistics on phase scintillations induced on the downlinks of Voyager 2 during a 1985 solar conjunction.

II. Measuring Phase Scintillations Induced on the Downlinks

Whenever a two-way coherent Doppler measurement is being performed, the downlink frequencies f_1 and f_2 are related to the uplink frequency f_0 by the transponding ratios G_1 and G_2 and the relative velocity of the spacecraft and the deep space station. In addition, there are noise terms that represent the phase scintillations picked up in transit through the solar corona. These frequency relationships may be expressed by

$$f_i = G_i f_0 (1 - 2\dot{\rho}/c) + \frac{G_i \nu_u^2}{f_0} + \frac{\nu_d^2}{f_0 G_i}; \qquad i = 1, 2 \quad (1)$$

The nonrelativistic Doppler shift has been indicated here; it is proportional to the ratio of the range rate $\dot{\rho}$ to the speed of light, c.

The parameter ν_u^2, in units of Hz^2, represents the level of phase scintillations induced on the uplink by the solar corona. It depends only on the physical properties of that part of the corona through which the uplink radio beam passes. It is an implicit function of time. By taking the integral of the square of the local plasma frequency along the uplink ray path, then taking a time derivative and dividing by twice the speed of light, one could calculate ν_u^2. Unfortunately, the local plasma frequency is generally not known. The parameter ν_d^2 is the downlink analog of ν_u^2.

Using Eq. (1), we can obtain an expression for ν_d^2:

$$\nu_d^2 = G_1 f_0 \left[1 - (G_1/G_2)^2\right]^{-1} [f_1 - (G_1/G_2) f_2] \qquad (2)$$

The phase scintillations induced on the downlinks contribute a frequency noise term that can be identified if the differential Dopper shift, $f_1 - (G_1/G_2) f_2$, is measured. Once identified, this frequency noise term is easily subtracted out of the Doppler record. At this point, the only error due to solar corona remaining in the Doppler record is the term involving ν_u^2. The Doppler error ϵ, in velocity units, is

$$\epsilon = \frac{c}{2 f_0^2} \langle \nu_u^2 \rangle_T \qquad (3)$$

The brackets $\langle \cdot \rangle_T$ indicate a time average over an integration time T.

The statistics of ν_u^2 and ν_d^2 are the same. This is in fact the key to being able to estimate the error ϵ. In the writing of standard deviations $\sigma(\cdot)$, then, the subscripts u and d may be ignored.

$$\sigma(\epsilon) = \frac{c}{2 f_0^2} \sigma(\langle \nu^2 \rangle_T) \qquad (4)$$

As suggested by Eq. (2), the required coronal statistics can be obtained by taking a statistical measure of the differential Doppler shift, viz.,

$$\sigma(\langle \nu^2 \rangle_T) = G_1 f_0 \left[1 - (G_1/G_2)^2\right]^{-1} \sigma(\langle f_1 - (G_1/G_2) f_2 \rangle_T) \qquad (5)$$

The required coronal statistics can also be obtained from a pair of one-way downlinks of different but related frequencies. This is, in fact, how the data appearing in this report were obtained. In this case, Eqs. (4) and (5) still hold, but some terms in Eq. (5) need proper interpretation. The term $G_1 f_0$ is the frequency of the first downlink, and the term G_1/G_2 is the ratio of the downlink frequencies.

III. Voyager 2 Results

During the 1985 solar conjunction for Voyager 2, differential Doppler data were collected from a pair of one-way downlinks. The downlinks originated with an ultrastable oscillator aboard the spacecraft, and the 64-m subnet was used for reception. Estimates were made of what the observed level of phase scintillations would do to a two-way coherent Doppler measurement. These estimates were obtained by applying the differential Doppler statistics to Eqs. (4) and (5). The frequency of the first downlink was approximately 2296 MHz. The ratio of the downlink frequencies, G_1/G_2, was 3/11.

The standard deviation of the two-way coherent Doppler error, as expressed in Eq. (4) and based on Voyager 2 differential Doppler measurements, has been compiled in Table 1 and plotted in Fig. 1. The statistics have been calculated for integration times ranging from 1 second to 1 minute. For the longer integration times, the error is less. Four example Sun-Earth-probe angles have been included. It must be understood that the properties of the solar corona are highly variable and that the results shown in Table 1 and Fig. 1 do not represent a characterization of the error as a function of Sun-Earth-probe angle. Such a characterization must be based on a larger set of data.

The Doppler error due to solar corona can be reduced by using a higher frequency on the uplink. For 7.2-GHz and 34.3-GHz uplinks, the errors shown in Table 1 and Fig. 1 are divided by 11.5 and 263, respectively. For some missions, this is an important advantage for the higher frequencies.

IV. Conclusions

This report has explained how phase scintillations induced on the downlinks can be measured. It has been shown that taking statistics on these measured phase scintillations leads to estimates of the Doppler error due to uplink phase scintillations. Typical estimates for the error have been calculated based on differential Doppler data collected during the 1985 Voyager 2 solar conjunction.

Table 1. Two-way coherent Doppler error for 2.1-GHz uplink and dual-frequency calibrated downlink*

Integration time, sec	Doppler error (mm/sec) for several Sun-Earth-probe angles			
	1.3°	1.7°	2.1°	2.5°
1	172.0	137.2	99.4	41.3
5	161.7	98.2	65.7	31.2
10	152.6	78.1	54.2	27.2
15	142.7	67.1	47.9	24.3
20	137.1	64.7	45.3	23.2
25	131.7	57.9	40.2	22.8
30	127.7	53.5	41.1	21.7
35	123.9	52.9	37.9	21.3
40	122.3	51.5	37.2	19.4
45	117.7	48.7	33.9	19.6
50	117.9	47.6	34.0	20.5
55	116.5	45.9	34.1	19.2
60	112.4	43.5	33.4	18.4

*These results are based on phase scintillation measurements made using the Voyager 2 spacecraft during its 1985 solar conjunction. The time and place at which the measurements were made are indicated below.

1.3°: 2030 to 2135 on 12-8-85 at Goldstone
1.7°: 0440 to 0640 on 12-12-85 at Canberra
2.1°: 0435 to 0545 on 12-8-85 at Canberra
2.5°: 2015 to 2120 on 12-12-85 at Goldstone

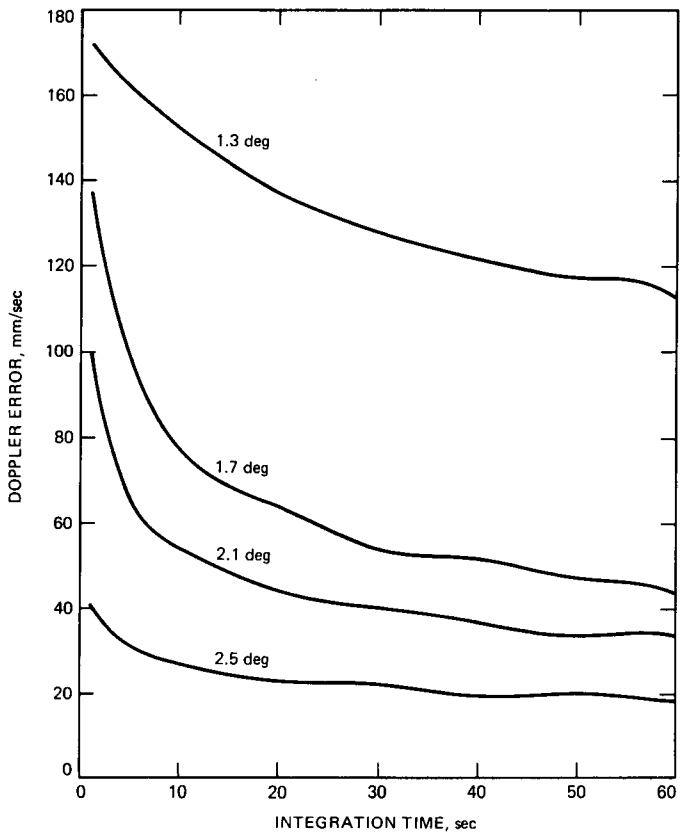


Fig. 1. Doppler error due to solar corona for various Sun-Earth-probe angles.

Dynamic Models for Simulation of the 70-M Antenna Axis Servos

R. E. Hill
Ground Antenna and Facilities Engineering Section

Dynamic models for the various functional elements of the 70-m antenna axis servos are described. The models representing the digital position controller, the linear and nonlinear properties of the physical hardware, and the dynamics of the flexible antenna structure are encoded in six major function blocks. The general modular structure of the function blocks facilitates their adaptation to a variety of dynamic simulation studies. Model parameter values were calculated from component specifications and design data. A simulation using the models to predict limit cycle behavior produced results in excellent agreement with field test data from the DSS 14 70-m antenna.

I. Introduction

The recent acquisition of a microcomputer workstation and a software package for modern control system analysis and simulation has enabled combining the linear dynamics of the antenna structure, the nonlinearities of friction, and the quantized, sampled data properties of the control algorithm in a single simulation model. The dynamic simulation program provides a building block approach to modeling complex systems such as the 70-m antenna axis servos. The models described here were developed from previous linear models [1], [2] in generalized building block forms. The form of the modularity was chosen to maintain generality and to facilitate future expansion or simplification of models for selected parts of the system.

The 70-m servo models described here are organized according to function into six major blocks which are shown (Fig. 1) interconnected to represent the overall axis-position servo. The overall model includes the salient properties of the control algorithm, the digital-to-analog converter, the electronic control amplifier, the hydraulic servovalve, the hydraulic motor with associated friction, the antenna structure and pedestal, and the axis encoder.

II. Dynamic Model Development

A. Position Loop Control Algorithm

The structure of the control algorithm is discussed in detail in [1]. The estimator and gain equations for the computer mode from [1] are repeated here to illustrate the development of the discrete system equations representing the algorithm block.

$$E(n+1) = \Phi E(n) + \Gamma U(n) + L Y_E(n+1) \qquad (1a)$$

for E_2 through E_6, and

$$E_1(n+1) = E_1(n) + E_2(n) - R(n) \qquad (1b)$$

$$U(n+1) = -\mathbf{K}E(n+1) + NR(n+1) \qquad (2)$$

where $E(n+1)$ is the state estimate column vector of [1], $[E_1\ E_2\ E_3 \ldots E_6]$ corresponding to time interval $n+1$. In [1] the $R(n)$ term in the equation for the integral error estimate, E_1, was omitted in error.

In practice, the antenna servo controller evaluates Eq. (1a) in two steps in order to minimize the computation delay within the servo loop, and implements Eq. (1b) to simplify the computation of E_1 and conserve computing time. Eq. (1b) is equivalent to Eq. (1a) for the special case where the first row of Φ is [1 1 0 0 0 0], the first elements of Γ and L are zero, and $R(n)$ is subtracted. Eq. (1a) can thus be extended to the more general form to include E_1

$$E(n+1) = \Phi E(n) + \Gamma U(n) + LY_E(n+1) + MR(n) \qquad (3)$$

where M is the column vector $[\Phi_{1,2}\ 0\ 0\ 0\ 0\ 0]$.

Next, substituting

$$Y_E(n+1) = Y(n+1) - H(\Phi E(n) + \Gamma U(n))$$

and

$$U(n) = -KE(n) + NR(n)$$

into Eq. (3) yields the familiar form for the estimator

$$E(n+1) = [\mathbf{I} - LH][\Phi - \Gamma K]E(n) + LY(n+1)$$
$$+ [\mathbf{I} - LH]\Gamma NR(n) + MR(n) \qquad (4)$$

With the use of Eq. (4), the controller output U and the position estimate E_2 can be expressed in a compact state space form for a system having three inputs and two outputs defined by Eqs. (2) and (3)

$$\begin{bmatrix} E(n+1) \\ U(n+1) \\ E_2(n+1) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \\ \mathbf{HA} & \mathbf{HB} \end{bmatrix} \begin{bmatrix} E(n) \\ Y(n+1) \\ R(n) \\ R(n+1) \end{bmatrix} \qquad (5)$$

where

$$\mathbf{A} = [(\mathbf{I} - LH)(\Phi - \Gamma K)]$$

$$\mathbf{B} = [L \quad (\mathbf{I} - LH)\Gamma N + M \quad 0]$$

$$\mathbf{C} = -\mathbf{KA}$$

$$\mathbf{D} = [-\mathbf{KB} + [0 \quad 0 \quad \mathrm{N}]]$$

Figure 2 illustrates the control algorithm block with inputs which are the encoder output, Y, and the current position command, R; its outputs are the rate command, U, and the encoder position estimate, E_1. The two blocks shown are standard building blocks provided by the simulation program. The previous rate command, $R(k)$, is derived from the unit delay. The discrete time equations represented by the **A**, **B**, **C**, and **D** matrices are evaluated at 50-msec intervals by the State Space block.

B. Electronic Component Models

The digital-to-analog converter in Fig. 1 is represented by a standard building block which quantizes a continuous input function.

The axis encoder is modeled by a standard quantizer building block from the simulation program catalog as shown in Fig. 1. The quantization level corresponds to 360 degrees per 2^{20} encoder increments.

The amplifier model block shown in Fig. 3 represents the dynamics of the hardware rate and acceleration limiters, the rate loop compensation networks, and the valve driver amplifier. Inputs are the rate command from the digital-to-analog converter and the motor rate, and the outputs are the current to the hydraulic valve and the voltage at the average tachometer circuit node. Parameter values are calculated from the component values in the schematic diagram,[1] and the properties of the valve coil. Details of these calculations are included in the Appendix.

C. Hydromechanical Component Models

The hydraulic valve model block is shown in Fig. 4 where the input is valve coil current and the output is no-load volumetric flow. The flow reduction due to the hydraulic pressure of the load is incorporated in the damping term of the motor model. This form of modeling simplifies the block interconnections by eliminating a pressure feedback path from the motor block to the valve block. The valve flow versus current is modeled by a simple dynamic lag followed by a hysteresis function and a deadzone. The deadzone corresponds to the underlapped condition of the valve spool face and the hystere-

[1] JPL Drawing J9479871D, Schematic Diagram, Analog D.W.B., (internal document), Jet Propulsion Laboratory, Pasadena, California, 1987.

sis results from friction associated with the spool motion. The resulting flow characteristic of this model, shown in Fig. 5, compares well with those from actual valve tests performed by the manufacturer.

Figure 6 shows the model of the hydraulic motor where the inputs are the no-load hydraulic flow and the load torque at the antenna bullgear. The outputs are the rate at the bullgear, the rate at the motor shaft, and the differential hydraulic pressure. The model incorporates the performance equations described in [2] and a more accurate model [3] for the friction associated with the motor and the gear reducer.

D. Flexible Structure Dynamics

A block diagram of the structure dynamic model is shown in Fig. 7. The model includes the gearbox stiffness, the residual structure inertia, axis damping and Coulomb friction, and state space models for the flexible modes of the structure and the antenna pedestal. The residual inertia represents that part of the total axis inertia that is not associated with any of the flexible modes. The inputs correspond to the hydraulic motor rate and an external disturbance torque. The first three outputs correspond to the angular positions at the bullgear, the axis encoder, and the intermediate reference assembly (IRA), respectively. The fourth and fifth outputs correspond to the axis reaction torque which is reflected back to the motor, and the encoder rate. It should be recognized that the bullgear and IRA positions are in absolute, or inertial, coordinates while the encoder position and rate are measured relative to the pedestal displacement.

The model is based on the equations of motion described in [2] with the addition of axis damping, friction, and pedestal dynamics. A block diagram representation of the structure flexible mode dynamics is shown in Fig. 8(a) where the input is the bullgear angle and the outputs are the IRA angle and the net reaction torque reflected to the bullgear. The corresponding state space equations are given by Eq. (6). The flexible alidade structure model, for the elevation axis, is shown in the block diagram of Fig. 8(b) and the corresponding state space equations are given by Eq. (7). The pedestal dynamic equations have the same form as those of the alidade structure used in the elevation model in Eq. (7). Numerical values of the structure parameter are listed in Table 1.

$$\begin{bmatrix} \dot{x} \\ \Theta_{IRA} \\ T_R \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} x \\ \Theta_B \end{bmatrix} \tag{6}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & & & & & \\ -\frac{K_1}{J_1} & 0 & & & & & \\ & & 0 & 1 & & & \\ & & -\frac{K_2}{J_2} & 0 & & & \\ & & & & \cdots & & \\ & & & & & 0 & 1 \\ & & & & & -\frac{K_N}{J_N} & 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 \\ \frac{K_1}{J_1} \\ 0 \\ \frac{K_2}{J_2} \\ \vdots \\ 0 \\ \frac{K_N}{J_N} \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} a_1 & 0 & a_2 & 0 & \cdots & a_N & 0 \\ -K_1 & 0 & -K_2 & 0 & \cdots & -K_N & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} a_0 \\ \sum_1^N K_i \end{bmatrix}$$

and where x represents the state vector and $\dot{x}$ its time derivative, and

$$N = \text{number of flexible modes}$$

$$a_0 \equiv 1 - \sum_1^N a_i \tag{7}$$

$$\begin{bmatrix} \dot{x} \\ \dot{\Theta}_A \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} x \\ T_{AR} \end{bmatrix}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{-K_{A1}}{J_{A1}} & 0 & \frac{K_{A1}}{J_{A1}} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{K_{A1}}{J_{A2}} & 0 & \frac{-(K_{A1}+K_{A2})}{J_{A2}} & 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 \\ \frac{1}{J_{A1}} \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0 \end{bmatrix}$$

with

$$K_{A1} \equiv J_{A1}\,\omega_{A1}^2$$

$$K_{A2} \equiv J_{A2}\,\omega_{A2}^2$$

E. Friction

The axis friction representation shown in Fig. 7 is subject to the limitations described in [3]. At this stage in the model development the total friction is lumped at the motor shaft as the actual distribution between motor shaft and antenna axis is unknown. Presumably, the distribution of friction to both sides of the gearbox stiffness could have a noticeable effect on dynamic behavior.

III. Model Tests and Applications to Simulations

The model was tested by using a feature of the simulation program that produces the linear system matrices representing the linearized model. The linear system matrices provide for a convenient cross check with other linear analysis methods thus assuring proper interconnect, feedback polarity, and parameter values. The Eigenvalues of the rate loop portion of the model of Fig. 1 were thus computed and results compared to three decimal place accuracy with those listed in [2]. Due to numerical condition deficiencies of the linear system matrix produced by the program, attempts to determine the transfer function zeros led to unreasonable results. This condition deficiency was overcome by scaling the relevant inertia and stiffness parameters according to the square of the gear ratio (see Fig. 6), and zeros in good agreement with [2] were thus obtained. The condition deficiency will be seen to affect only the derived linear system matrix and will not impair accuracy of simulation runs. Following this cross check process the model was restored to the unnormalized form shown in Figs. 6 and 7.

A simulation of position loop limit cycle behavior was performed to compare the performance of the overall model including the new friction model with results from tests on the actual antenna. Results from the simulation shown in Fig. 9 are remarkably similar to those from the antenna test of Fig. 10. In both the simulation and the hardware test, the limit cycle was initiated by a small (3 encoder bits) position input step change. The simulation was performed at a stage prior to the development of the latest models for the discrete time control algorithm, the amplifier, and the hydraulic valve, so linear dynamic equivalents were substituted. For the algorithm, the proportional-plus-integral-plus-derivative (PID) linear feedback equivalent was used. The model structure and parameter values were otherwise identical to those described above.

IV. Summary and Conclusions

Modular dynamic models for the 70-m antenna axis servos have been described. Numerical cross checks of Eigenvalues and transfer function zeros indicate a consistency between the model and previous linear analysis methods. Comparisons of model-based simulation results with actual field test results indicate excellent modeling accuracy.

The small discrepancies between the simulation and field test results are most likely the result of differences between the modeled and actual friction parameter values, and the presence of a small (150 psi) bias torque effect in the actual antenna. In future work the model and the simulation program should be extremely useful in two ways: (1) as an adjunct to the design of more robust systems, and (2) as a hardware diagnostic aid that relates specific hardware out-of-tolerance conditions to abnormal field test measurements.

References

[1] R. E. Hill, "A Modern Control Theory Based Algorithm for Control of the NASA/JPL 70-Meter Antenna Axis Servos," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 285–294, November 15, 1987.

[2] R. E. Hill, "A New State Space Model for the NASA/JPL 70-Meter Antenna Servo Controls," *TDA Progress Report 42-91*, vol. July–September 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 247–264, November 15, 1987.

[3] R. E. Hill, "A New Algorithm for Modeling Friction in Dynamic Mechanical Systems," *TDA Progress Report 42-95*, this issue.

Table 1. Parameters for 70-m AZ and EL axis servos

FLEXIBLE MODES[a]

Stiffness, K, ft-lb/rad

K AZ = [2.238 4.516 3.651 1.233 0.564]^T · 1.E09

K EL = [22.61 6.564 1.566 3.406 0.699]^T · 1.E09

Squared natural frequencies, ω, (rad/sec)^2

ω AZ = [63.47 69.09 97.81 189.0 285.68]

ω EL = [218.77 313.04 411.12 476.72 656.85]

Transformation coefficients, a dimensionless

a AZ = [0.1331 0.2767 0.1169 0.0099 0.0376]^T

a EL = [0.2939 0.0977 0.0266 0.0309 0.0050]^T

ALIDADE STRUCTURE, ELEVATION AXIS ONLY

Inertia moments, ft-lb-sec^2, referred to the EL axis

JA1 = 1.197E07

JA2 = 1.098E08

Squared natural frequencies, ω, (rad/sec)^2

ωA1 = 4600

ωA2 = 1026

PEDESTAL STRUCTURE, AZIMUTH AXIS ONLY

Inertia moments, ft-lb-sec^2, referred to the AZ-axis

JP1 = 2.105E08

JP2 = 1.74E08

Squared natural frequencies, ω, (rad/sec)^2

ω P1 = 3.093E04

ω P2 = 4.063E04

BULLGEAR RESIDUAL INERTIA, J_B, ft-lb-sec^2

J_B AZ = 1.4965E08

J_B EL = 8.7989E07

GEARBOX STIFFNESS, K_G, ft-lb/rad

K_G AZ = 2.1654E11

K_G EL = 3.0759E11

JBINV AZ = 1/J_B AZ

JBINV EL = 1/J_B EL

AXIS DAMPING, ft-lb/radian/sec, (equiv 1000 psi/deg/sec)

2.13E08

GEAR RATIO (MOTOR:AXIS)

= 28730 (BOTH AXES)

MOTOR DISPLACEMENT (total 4 motors), V, in.^{3}/rad

V = 1.528

VALVE GAIN, K_V, in.^{3}/s/mA

K_V = 38.3

HYD COMPRESSIBILITY, C, in.^{3}/psi

C = .00314

VALVE DAMPING RATIO, D/C, sec^{-1}

D = 0.60

MOTOR INERTIA (total 4 motors), J_M, ft-lb-sec^2

J_M AZ = 1.00

J_M EL = 0.664

FRICTION, INERTIA, SAMPLE INTERVALS FOR FRICTION BLOCK, FRICTION EQUIVALENT TO 350 psi OF ΔP, STATIC FRICTION EQUIVALENT TO 420 psi

FRC = 44.57

STC = 53.48

JMFT = [J_M AZ FRC 0.005]

FRCS = [FRC STC]

[a]Calculated from [2], Table 5 using $K_i = (J_i/J_b)\,\omega_i^2\,J_b\,N^2$ with gear ratio N = 28730 for either axis.

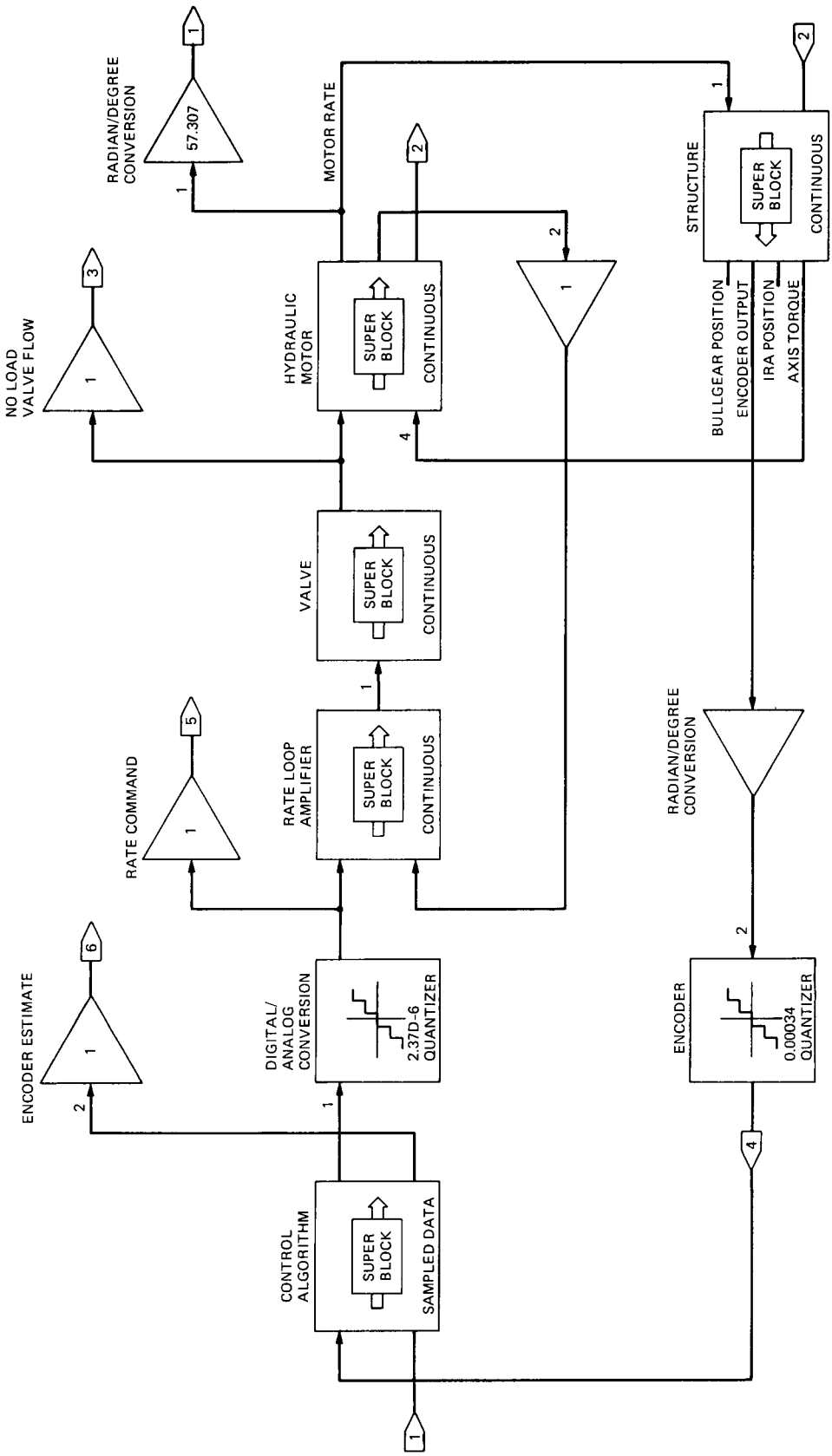


Fig. 1. Azimuth/Elevation servo simulation diagram.

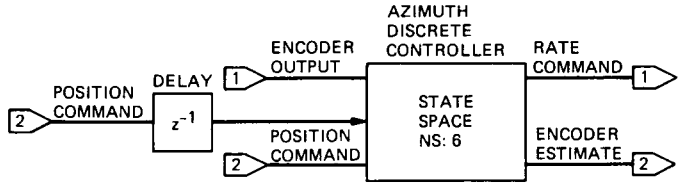


Fig. 2. Control algorithm simulation diagram.

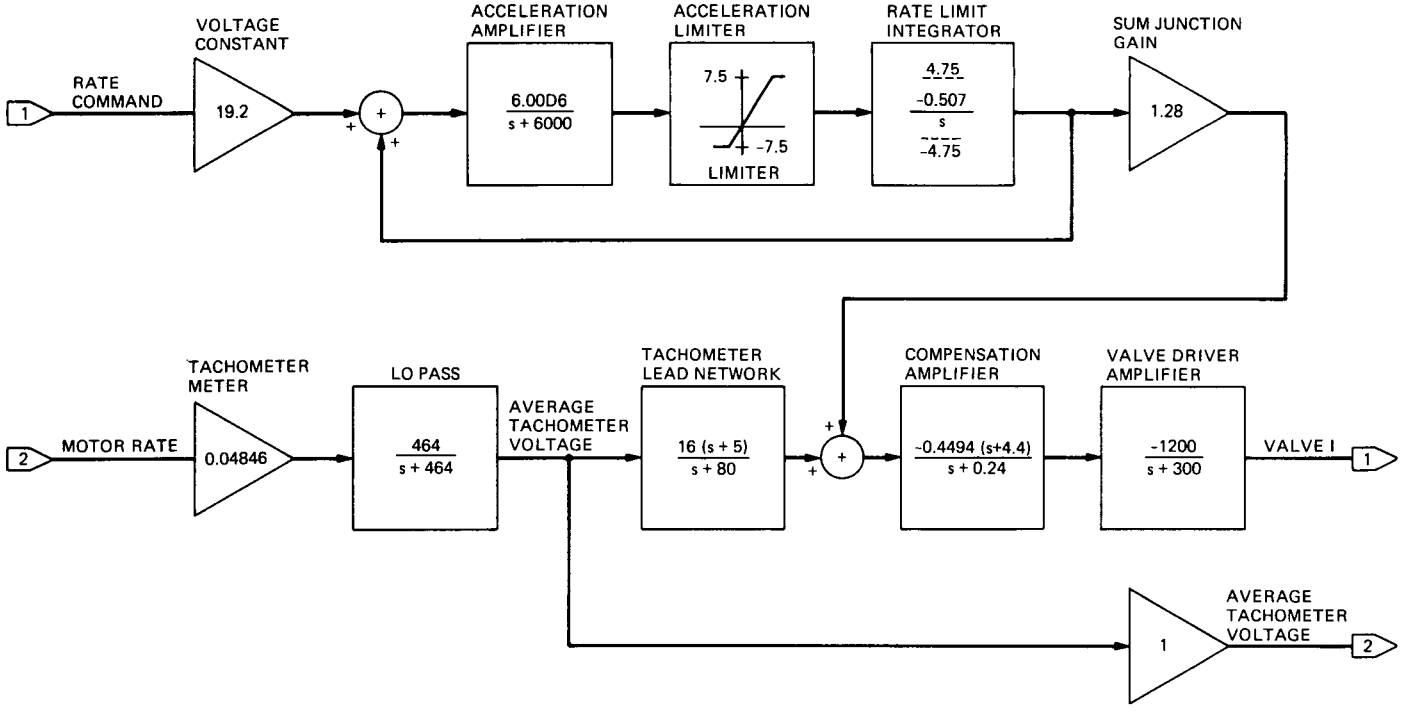


Fig. 3. Rate loop amplifier simulation diagram.

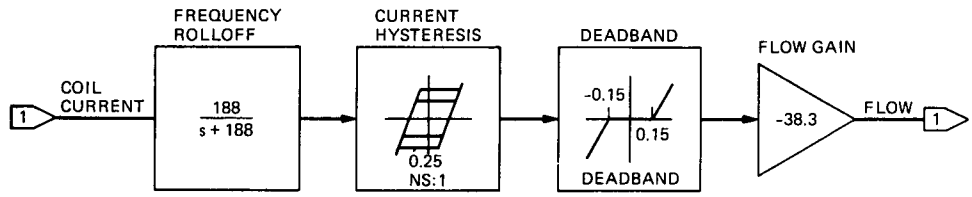


Fig. 4. Hydraulic valve simulation diagram.

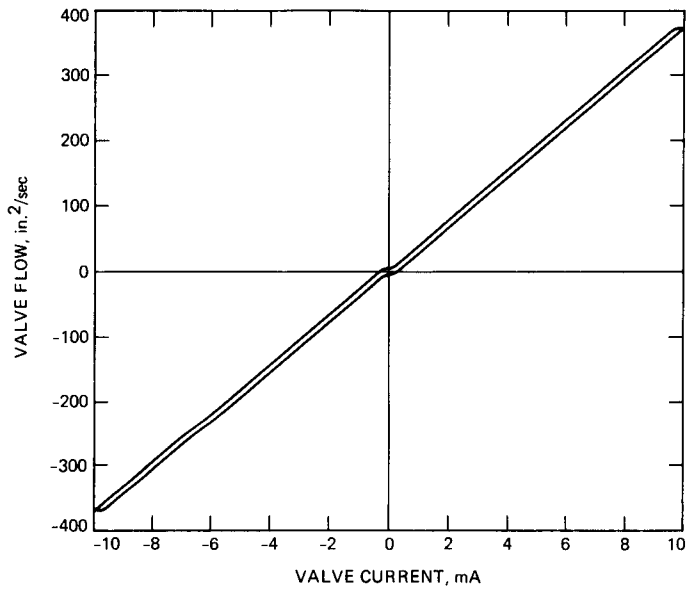


Fig. 5. Hydraulic valve flow characteristic.

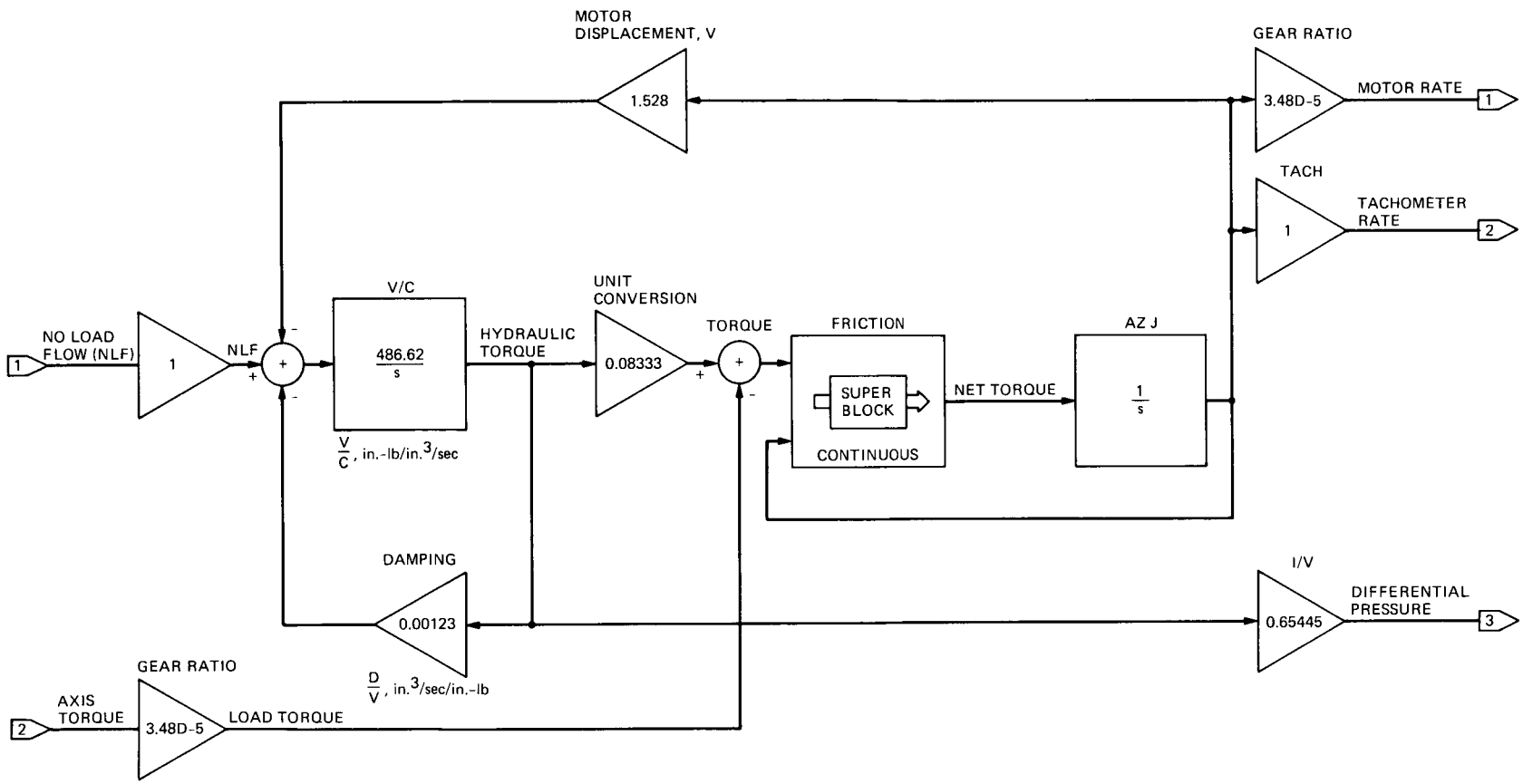


Fig. 6. Azimuth motor simulation diagram.

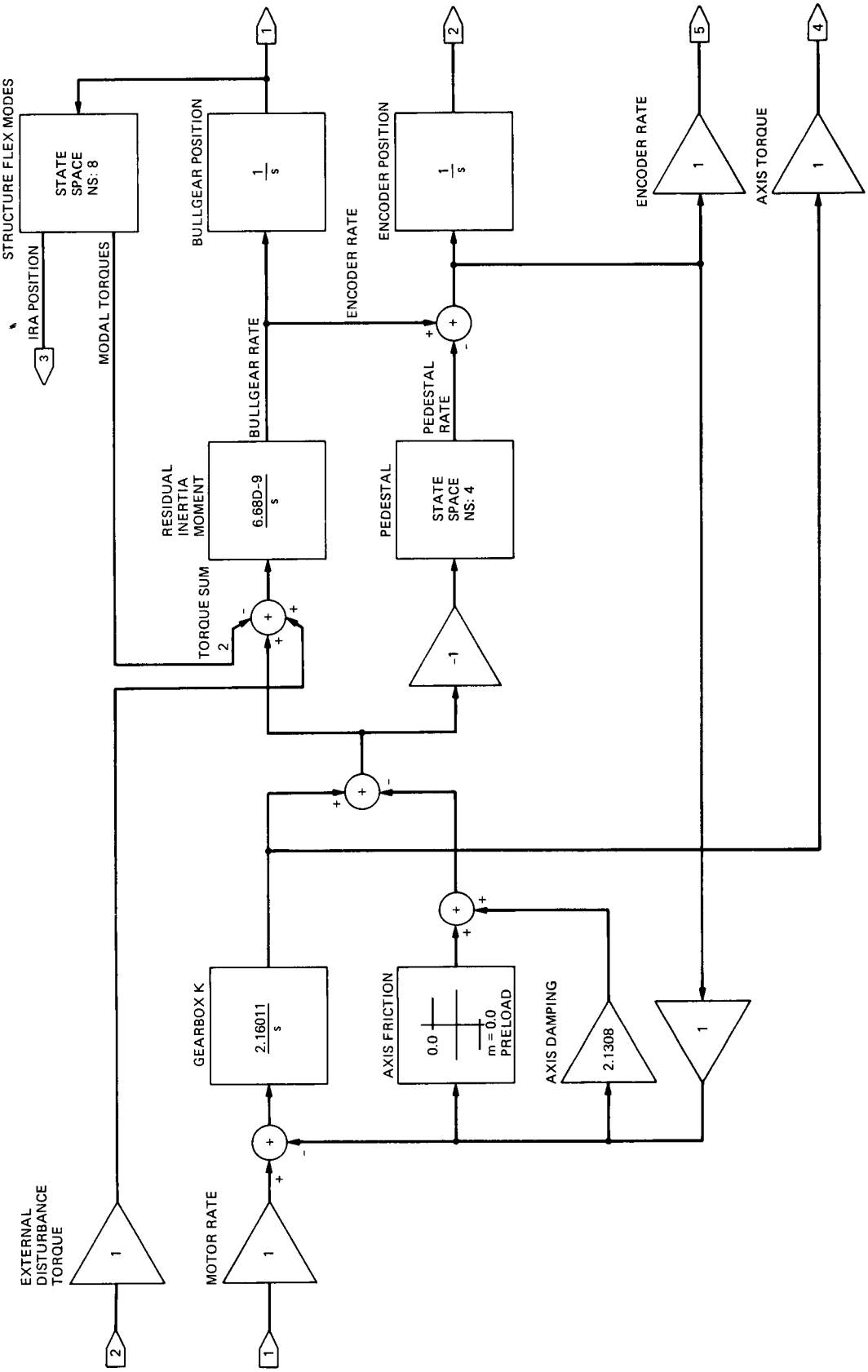


Fig. 7. Azimuth structure simulation diagram.

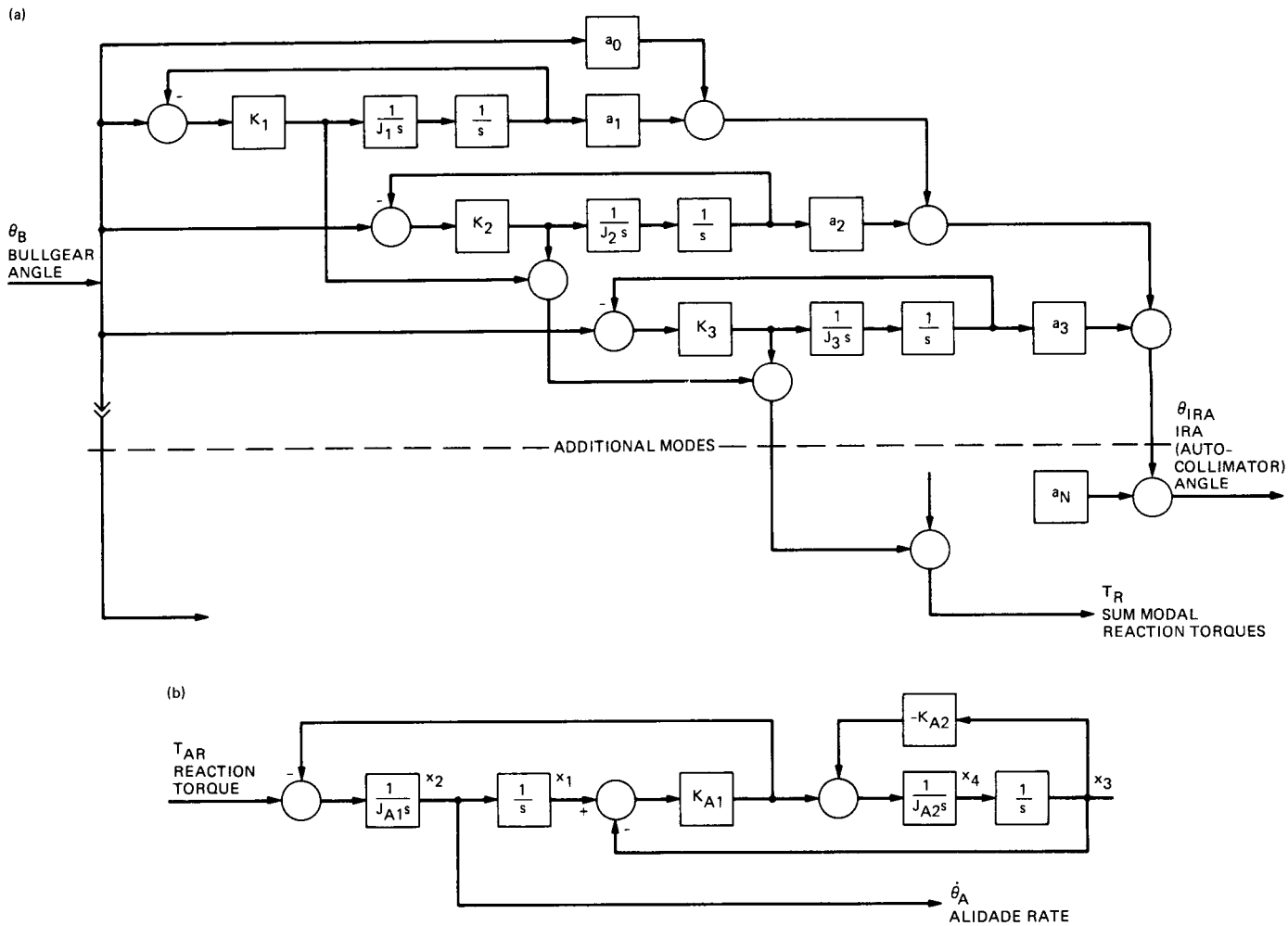


Fig. 8. 70-m simulation block diagram: (a) flexible modes and (b) alidade structure.

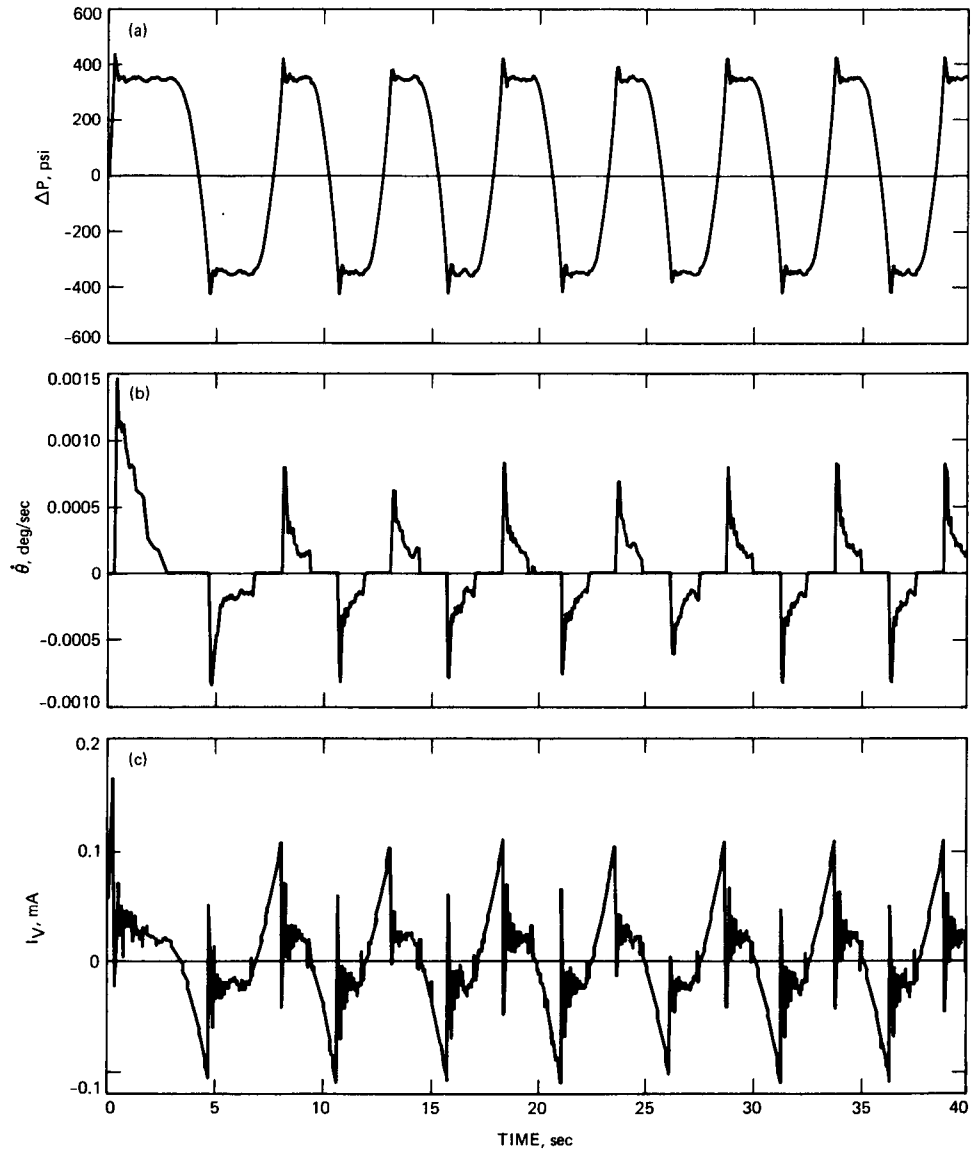


Fig. 9. Simulation test results for an azimuth limit cycling condition: (a) differential pressure; (b) axis rate; (c) hydraulic valve current; (d) rate command; and (e) encoder angle.

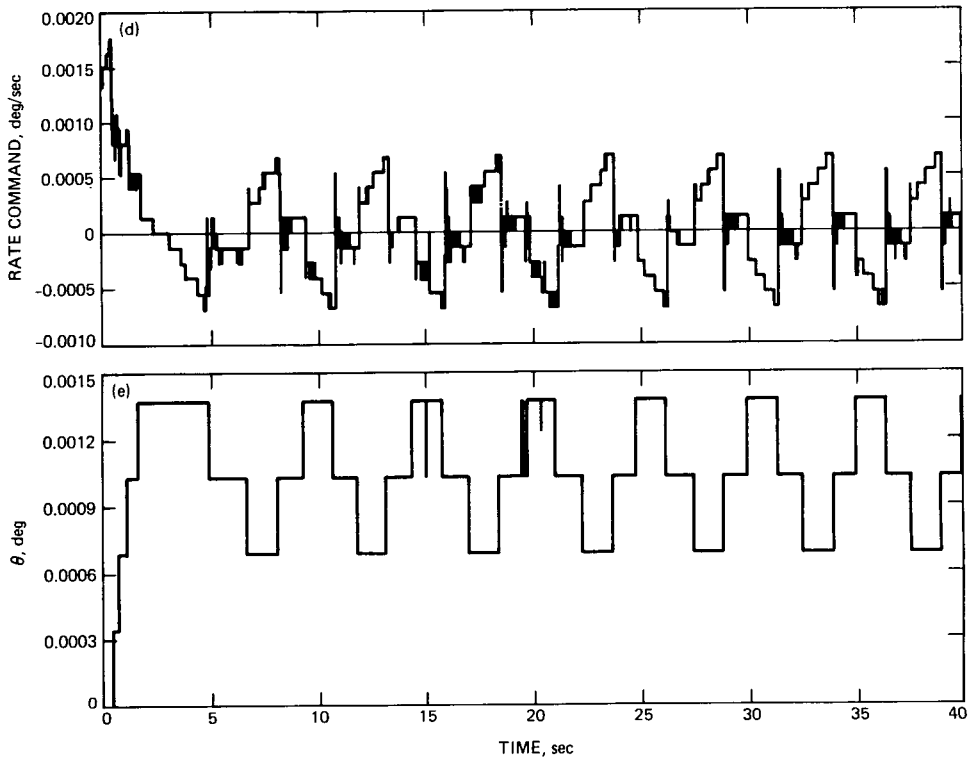


Fig. 9 (contd).

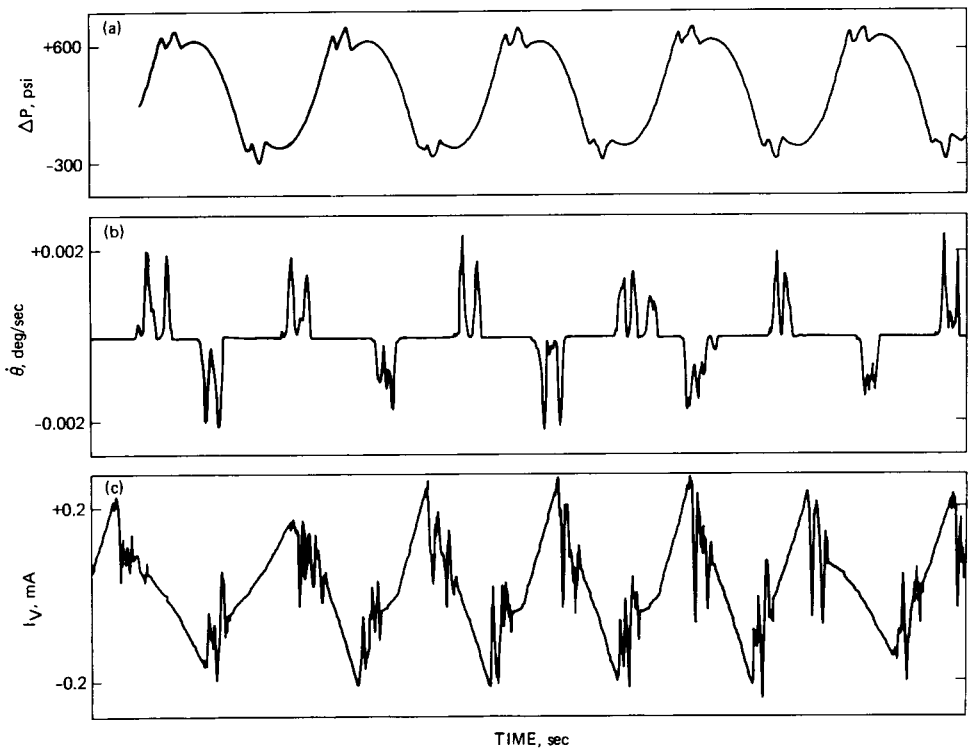


Fig. 10. Azimuth limit cycle test results from the DSS14 70-m antenna: (a) differential pressure; (b) axis rate; and (c) hydraulic valve current.

Appendix

Derivation of Electronic Circuit Transfer Functions

This appendix describes the analytical methods and detailed calculations used to derive the dynamic transfer functions of the electronic circuit portions of the 70-m axis servo loops. A number of simplifying approximations employed in the original design and network synthesis process are also described. The transfer functions are described in terms of a group of function blocks representing simplified equivalent circuits of portions of the analog circuit board and the external ratemeter. The basic blocks consist of a tachometer combining network and filter, a tachometer lead network, a compensation amplifier, a valve driver amplifier, and a rate and acceleration limiter.

I. Tachometer Combining Network and Filter

The tachometer network and filter properties are derived from the schematic diagram (see Footnote 1) and the external ratemeter components. The applicable part of the schematic and the ratemeter is shown in Fig. A-1(a). The four tachometer voltage divider networks, R34, R35, C17, etc. are omitted from Fig. A-1(a) because by virtue of their high impedance relative to the tachometer source resistances, these networks have negligible effect on the overall transfer functions. The V_{AT} symbol designates the "average tachometer" circuit node voltage, a significant node because it serves as a calibration reference for transfer function calculations. The ratemeter adjust potentiometer is adjusted in the field to compensate for disconnection of one or more tachometers, and also for scale factor variations among individual tachometers. Because this adjustment is in a shunt circuit path, it simultaneously corrects the rate loop gain and the voltage at V_{AT}, with adjustment of the high and low rate ranges of the meter circuit.

The voltage scale factor at V_{AT} to satisfy the calibration condition is calculated from the known full scale meter current, (50 μA), the meter circuit resistance consisting of fixed resistor R60 and the meter coil resistance R_m, and the full scale rate (0.25 deg/sec). The average tachometer circuit node voltage constant is thus

$$V_{AT} = \frac{(\text{R60} + R_m)\, I_{FSM}}{Rate_{FS}}$$

With R60 = 121 kΩ, R_m = 660 Ω

$$V_{AT} = \frac{(121 + 0.660)\ .05}{0.25}$$

$$= 24.33 \text{ volts/degree/sec axis rate}$$

$$= 0.04852 \text{ volts/rad/sec of motor rate}$$

This value is as accurate as the combined accuracy of the R60 resistor (1 percent), the meter movement calibration (2 percent), and the field calibration process (estimated 3 percent) which is sufficient for the present purpose. A small error (4.9 percent) in the low rate range calibration results from the difference of the ratio of (R60 + 660)/(R61 + 660) from the desired 10:1. This error could be diminished by padding R61 with 220 kΩ.

The network of Fig. A-1(a) is reduced to the equivalent of Fig. A-1(b) by replacing the four inputs with the single equivalent voltage source V_T, replacing the R54, R57, R58, R59, R62 combination with the series resistor R_{ST}, and replacing the meter circuit with the shunt resistor R_M. Calculation of the shunt resistor R_M requires a knowledge of the setting of the 100 K ratemeter adjustment potentiometer. This is accomplished indirectly through use of the known value of V_{AT} derived from the ratemeter adjustment criteria described above and the tachometer scale factors. Thus, representing the sum of the shunt conductances by G_{SS} and with $G_{ST} = R_{ST}^{-1}$:

$$G_{SS} = G_{ST}\left(\frac{V_T}{V_{AT}} - 1\right)$$

which leads to

$$R_M^{-1} = R_{ST}^{-1}\left(\frac{V_T}{V_{AT}} - 1\right) - (\text{R63} + \text{R64})^{-1} - (\text{R65} + \text{R66})^{-1}$$

where

$$R_{ST} = (\text{R56}^{-1} + \text{R57}^{-1} + \text{R58}^{-1} + \text{R59}^{-1})^{-1} + \text{R62}$$

With V_T = 187.5 and 222 volts/deg/sec for azimuth and elevation, respectively, and with

$$\text{R56} = \text{R57} = \text{R58} = \text{R59} = 8.25 \text{ k}\Omega$$

$$\text{R62} = 56.2 \text{ k}\Omega$$

$$\text{R63} = 40.0 \text{ k}\Omega$$

$$R64 = 51.1\ k\Omega$$

$$R65 = 909\ k\Omega$$

$$R66 = 51.1\ k\Omega$$

The equivalent meter circuit resistance becomes

$$R_M = 9.73\ k\Omega \text{ for azimuth}$$

$$R_M = 7.847\ k\Omega \text{ for elevation}$$

The 56.2 K resistor R14 provides part of the additional shunt conductance required for elevation.

The network poles and zero are calculated from general expressions derived from the circuit loop equations using Cramer's method with expansion of the determinants. The computation is simplified to a two-pole, single-zero determination by partitioning the circuit to delete C42, R65, and R66 from the circuit. This approximation is justified since the response zeros correspond to the parallel resonance of R65 and C42 and to the series resonance of the R63, R64, C40 branch and are thus unaffected by the partitioning. The effect of partitioning on the response poles is small because of the large ratio of R66 to R_M. The exact location of the tachometer filter poles is also of secondary importance to control dynamics analysis. The circuit equations may be derived from any consistent set of loop currents or node voltages. A typical set of loop current equations is

$$\begin{bmatrix} R_1 + R63 + R64 & -R64 & -R63 \\ -R64 & R64 + (sC40)^{-1} & -(sC40)^{-1} \\ -R63 & -(sC40)^{-1} & R63 + (sC40)^{-1} + (sC41)^{-1} \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix} = \begin{bmatrix} V_1 \\ 0 \\ 0 \end{bmatrix}$$

where $R_1 = (R_{ST}^{-1} + R_M^{-1})^{-1}$ and $V_1 = V_T R_M (R_{ST} + R_M)^{-1}$. Expansion of the determinant yields the following quadratic equation in s:

$$s^2 + [(R_1^{-1} + R63^{-1})\,C41^{-1} + (R63^{-1} + R64^{-1})\,C40^{-1}]s$$

$$+ (R_1 + R63 + R64)\,(R_1\ R63\ R64\ C41\ C40)^{-1} = 0$$

the roots of which are the circuit poles. Expansion of the numerator cofactor yields the single zero

$$Z1 = -(R63^{-1} + R64^{-1})\,C40^{-1}$$

Numerical evaluation for the azimuth axis results in

$$Z1 = -297$$

$$P1 = -258$$

$$P2 = -1006$$

The transfer function relating the average tachometer node voltage to the motor rate in radians/sec becomes

$$\frac{V_{AT}}{\dot{\Theta}_M} = \frac{-0.4852\ P2}{(s - P2)} = \frac{488.1}{(s + 1006)}$$

Because of the near cancellation of Z1 and P1, they are omitted from the model and the circuit is approximated by the single pole, P2. As an alternative to the quadratic solution above, a convenient approximation for the poles may be employed whereby

$$P1 = -(R63^{-1} + R64^{-1})\,C40^{-1}$$

$$P2 = -(R_1^{-1} + R63^{-1})\,C41^{-1}$$

This approximation yields -297 and -967 for P1 and P2.

II. Tachometer Lead Network

The calculation of the tachometer lead network pole and zero is simplified by partitioning the circuit and replacing the tachometer network by an equivalent source resistance, R_S. The lead pole and zero frequencies thus become

$$PL1 = -(R_S + R65 + R66)\,[(R_S + R66)\,R65\,C42]^{-1}$$

$$ZL1 = -(R65\,C42)^{-1}$$

where $R_S = [R_{ST}^{-1} + R_M^{-1} + (R63 + R64)^{-1}]^{-1}$.

Numerical results are

$$PL1 = -82.38 \text{ for azimuth}$$

$$PL1 = -84.01 \text{ for elevation}$$

$$ZL1 = -5.00 \text{ for both axes}$$

III. Rate Loop Compensation Amplifier

The simplified schematic circuit diagram of the rate loop amplifier and the tachometer lead network is shown in Fig. A-2 where R_S is the network source resistance discussed earlier. Using the infinite summing junction gain approximation, the high frequency voltage gain of the circuit is the ratio of feedback to input impedances where the reactances of the capacitors are zero. This high frequency gain is subsequently cascaded with the zero/pole ratios of the input and feedback networks to derive an overall transfer function. Thus

$$\frac{V_{RA}}{V_{AT}} = \frac{R53}{(R_S + R66)}$$

The circuit pole and zero corresponding to the feedback network are derived using Cramers method of solution of the circuit loop equations applicable to the infinite gain approximation. Solution of the equations yields for the pole and zero frequencies

$$PR1 = -C31^{-1}\left[R50 + R52\left(1 + \frac{R50}{R51}\right)\right]^{-1}$$

$$ZR1 = -C31^{-1}\,[R53^{-1} + (R52 + R50\,R51)(R50 + R51)]^{-1}$$

Numerical results are

$$PR1 = -0.238\ s^{-1}$$

$$ZR1 = -4.47\ s^{-1}$$

The voltage transfer function of the tachometer lead network and the loop compensation amplifier thus becomes

$$\frac{V_{RA}}{V_{AT}} = \frac{-R53\,(s - ZL1)\,(s - ZR1)}{(R_S + R66)\,(s - PL1)\,(s - PR1)}$$

$$= \frac{-7.61\,(s + 5.0)\,(s + 4.47)}{(s + 0.238)\,(s + 82.4)}$$

IV. Valve Driver Amplifier

The equivalent circuit of the valve driver amplifier is shown in Fig. A-3 where V_{AR} is the input voltage from the rate amplifier and I_V is the output current in the hydraulic valve load. The Q1, Q2 complementary transistor emitter followers are represented by the unity voltage gain block. Using the infinite gain approximation for the operational amplifier, neglecting the gain-bandwidth product, and including the phase inversion of the op-amp, the circuit transconductance becomes

$$\frac{I_V}{V_{RA}} = \frac{-1}{(R13\ R43\ C18)\,(s - PV1)} = \frac{-1217}{(s + 303)}$$

where $PV1 = (R36\ C18)^{-1} = -303$.

V. Rate and Acceleration Limiters

The equivalent circuits of the rate and acceleration limiters are shown in Fig. A-4 where V_{RC} represents the rate input command voltage from the external digital to analog converter in the antenna servo controller and V_{RL} represents the limiter output.

The voltage scale factor at V_{RC} is derived from the equilibrium condition where a V_{RC} input command is opposed by a tachometer input, V_{AT} (see Figs. A-1 and A-2) such that their difference results in a value of V_{RA} sufficient to produce the desired rate. This difference is inversely proportional to the negative loop gain of the rate loop at DC, K_{DC}. Thus, using R38/R23 as an approximation to the DC transfer function of the limiter circuit,

$$\frac{V_{AT}}{(R65 + R66)}\,\frac{-V_{RC}\,R38}{(R23\ R15)} = \frac{-K_{DC}\,V_{AT}}{(R65 + R66)}$$

from which

$$V_{RC} = \frac{V_{AT}\,(1 + K_{DC}^{-1})\left(\frac{R15\ R23}{R38}\right)}{(R65 + R66)}$$

with the loop gain $K_{DC} = 40$ and $V_{RC} = 19.48$ volts/deg/sec.

Normal variations of the rate loop gain about the typical value of 40 will result in errors of negligible proportion relative to the uncertainties of tachometer and hydraulic valve gains.

The acceleration amplifier U7 with feedback network R32 and R33 is equivalent to a voltage gain with a single real

pole resulting from the gain-bandwidth product of the LF356 operational amplifier. Neglecting terms in the operational amplifier gain, 1/A, the closed-loop voltage gain and pole frequency become

$$\frac{V_{AL}}{V_{RC}} = 1 + \frac{\text{R33}}{\text{R32}}$$

$$P_{AL} = -2\pi \text{GBW} \frac{\text{R32}}{(\text{R32} + \text{R33})}$$

With R32 = 1.0 K, R33 = 1.0 M, and GBW = 1.0 MHz

$$\frac{V_{AL}}{V_{RC}} = 1001$$

$$P_{AL} = -6280$$

The acceleration voltage limiter threshold, V_{ALT}, is the sum of the forward and zener voltages (0.7 + 6.8) of the 1N5526 Zener diodes, CR1 and CR2.

The action of the rate limiter is represented in Fig. A-4 by the limiter block in parallel with capacitor C24. Typical settings of the adjustable limiter correspond to a rate limit of 0.25 deg/sec which, using 19.48 volts/deg/sec, equals a 4.87 volt limit at V_{RL}. The actual setting of the acceleration limit adjustment R37 can be calculated from the nominal component values and an assumed adjustment to 0.20 degrees/sec^2. Using the equation for the integrator transfer function

$$V_{AL} = R_{IN}\, \text{C24} \frac{d}{dt} V_{RL}$$

where R_{IN} is the actual input resistance, R39 plus the adjusted value of R37

$$R_{IN} = V_{AL} \left(\text{C24} \frac{d}{dt} V_{RL}\right)^{-1}$$

Substituting the acceleration limit threshold voltage, 7.50, for V_{AL}, 19.48 volts/sec per deg/sec multiplied by 0.2 deg/sec^2 for $d/dt\ V_{RL}$,

$$R_{IN} = 7.50\,(19.48 \cdot 0.2\ \text{C24})^{-1} = 1.925\ \text{M}\Omega$$

from which the center value of R36 is 325 kΩ for an acceleration limit of 0.2 deg/sec^2.

Neglecting the high frequency pole, P_{AL}, the transfer function from the rate command to V_{AL} becomes

$$\frac{V_{AL}}{\dot{\Theta}_{RC}} = \frac{19.48\left(\frac{\text{R38}}{(\text{R23} + \text{R38})}\right)\left(1 + \frac{\text{R33}}{\text{R32}}\right)}{1 + \left(\frac{\text{R38}}{(\text{R23} + \text{R38})}\right)\left(1 + \frac{\text{R33}}{\text{R32}}\right)(R_{IN}\,\text{C24}\,s)^{-1}}$$

which, due to the high value of loop gain, can be approximated by

$$\frac{V_{AL}}{\dot{\Theta}_{RC}} = 19.48\, R_{IN}\, \text{C24}\, s = 38.50\ \text{volts/degree/sec}^2$$

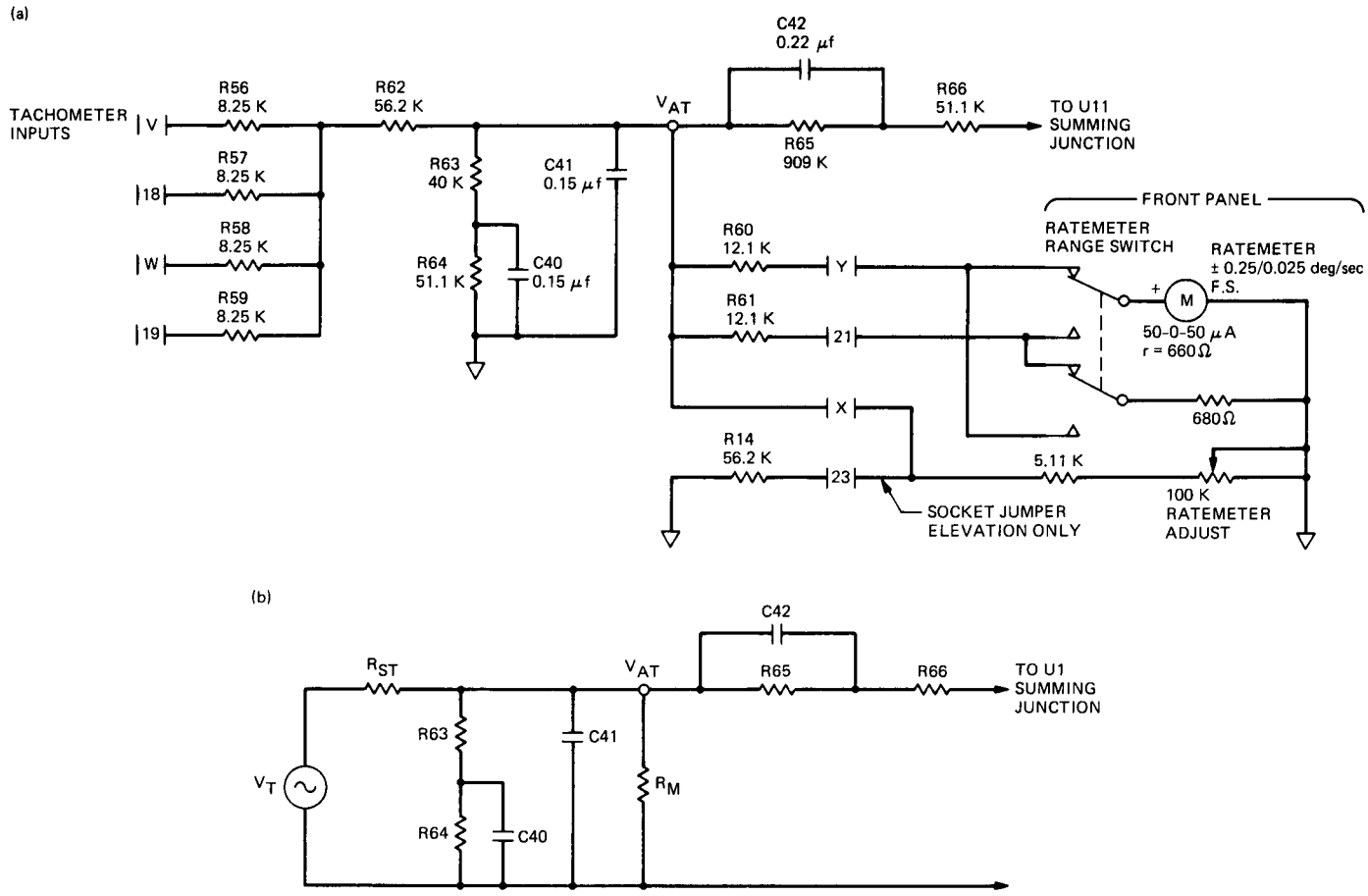


Fig. A-1. Tachometer combining network: (a) simplified schematic diagram and (b) equivalent circuit.

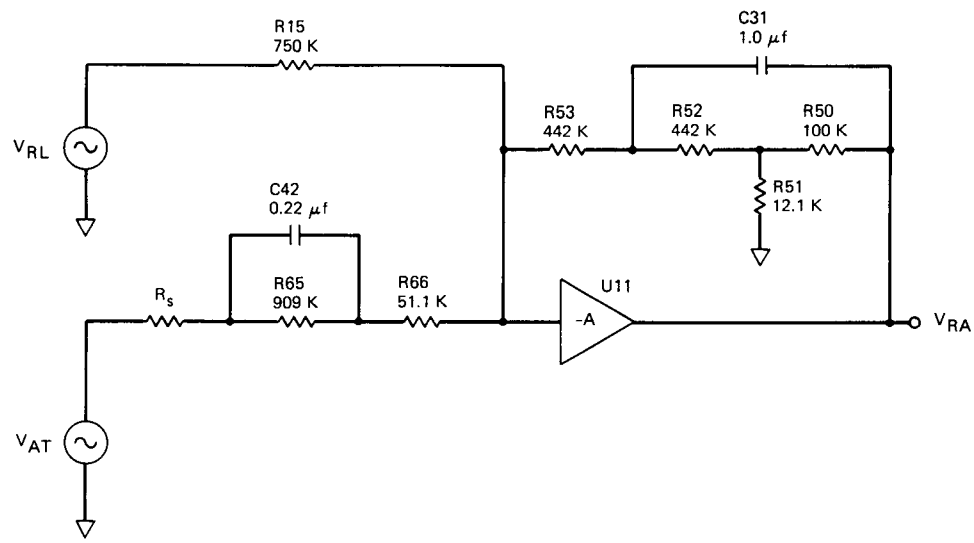


Fig. A-2. Rate amplifier equivalent circuit.

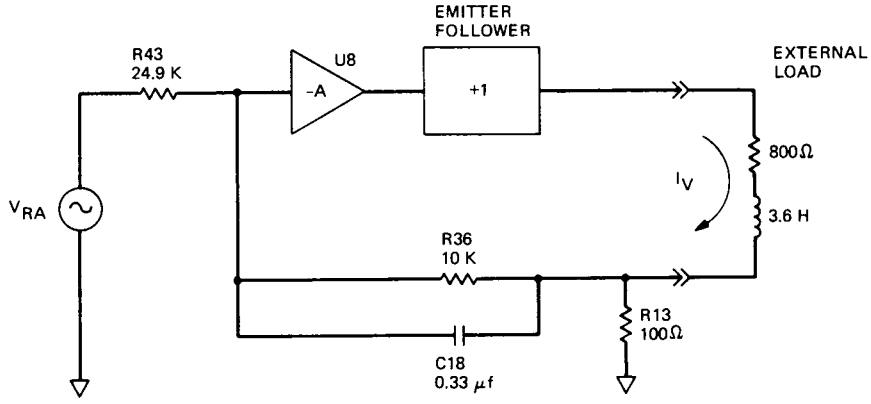


Fig. A-3. Valve driver amplifier equivalent circuit.

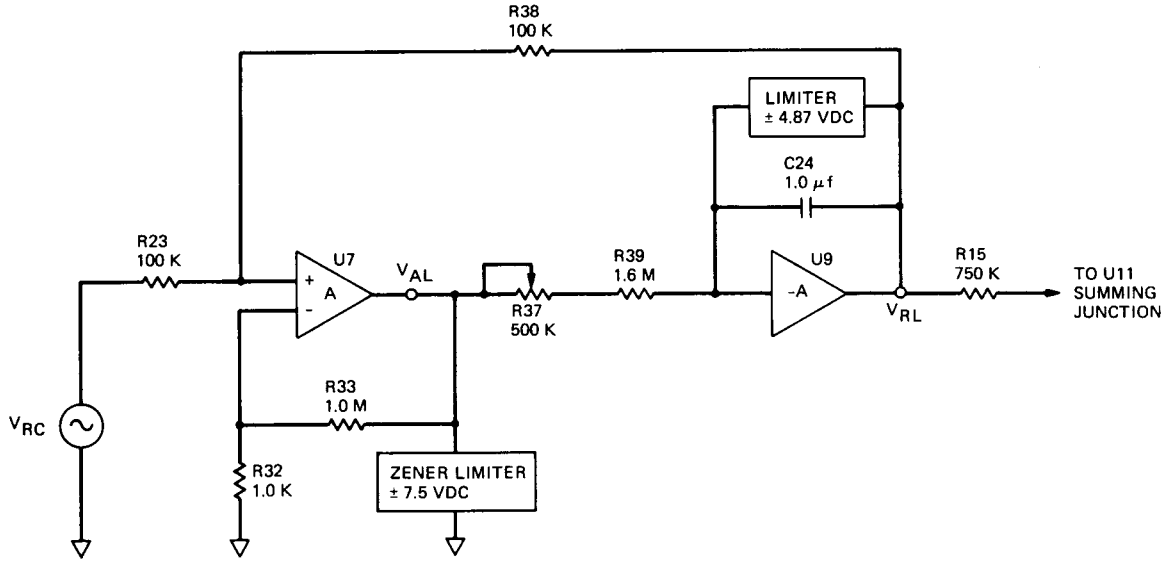


Fig. A-4. Rate and acceleration limiter equivalent circuit.

A New Algorithm for Modeling Friction in Dynamic Mechanical Systems

R. E. Hill
Ground Antenna and Facilities Engineering Section

A new method of modeling friction forces that impede the motion of parts of dynamic mechanical systems is described. Conventional methods in which the friction effect is assumed a constant force, or torque, in a direction opposite to the relative motion, are applicable only to those cases where applied forces are large in comparison to the friction, and where there is little interest in system behavior close to the times of transitions through zero velocity. This article describes a new algorithm that provides accurate determination of friction forces over a wide range of applied force and velocity conditions. The method avoids the simulation errors resulting from a finite integration interval used in connection with a conventional friction model, as is the case in many digital computer-based simulations. The new algorithm incorporates a predictive calculation based on initial conditions of motion, externally applied forces, inertia, and integration step size. The predictive calculation in connection with an external integration process provides an accurate determination of both static and Coulomb friction forces and resulting motions in dynamic simulations. Accuracy of the results is improved over that obtained with conventional methods and a relatively large integration step size is permitted. A function block for incorporation in a specific simulation program is described. The general form of the algorithm facilitates implementation with various programming languages such as Fortran or C, as well as with other simulation programs.

I. Introduction

Recent interest in certain limit cycle oscillatory modes of operation of the 70-m antenna at DSS 14 has intensified the need for dynamic analysis and simulation of the axis servos. Limit cycle oscillations of phyiscal positioning systems, such as the antenna axis servos, result from nonlinearities associated with the position sensors and the control actuation devices. To support limit cycle investigations it is necessary to model and simulate all the identified nonlinearities in the system. Because of the large magnitude of axis friction, which equals roughly 20 percent of the maximum available control effort, accurate modeling of friction effects is of critical importance.

The basic physical laws of friction are discussed in numerous textbooks on mechanics and are briefly summarized here. First, the friction force between two bodies lies in the tangent plane of the contact point between the bodies. In the absence of relative motion, its magnitude is less than or equal to the

product of the normal force between the bodies and a constant coefficient of static friction. Relative tangent plane motion between the bodies cannot commence until an externally applied force exceeds the maximum magnitude of the friction force. Second, the Coulomb (as opposed to viscous) friction in the presence of relative motion between the two bodies, the friction force equals the product of the normal force and a coefficient of friction which may be different from the static coefficient and is in a direction opposite to the relative motion.

In those applications where the normal force is constant and static friction forces are of no great concern, the friction force can be modeled by a constant force directed opposite the relative motion. This conventional model combined with an inertia is illustrated in transfer function form in Fig. 1 where motion is restricted to a single coordinate. The nonlinear function block in Fig. 1 has three possible outputs: a unit amplitude with algebraic sign the same as that of the velocity; $\dot{x}$, (when $\dot{x}$ is nonzero); and zero output for zero $\dot{x}$. The constant of multiplication, F_c, in the constant block is the product of the normal force and the friction coefficient. The net input force to the integration block is thus equal to the difference between the applied force and the friction force.

By inference, the behavior of the model of Fig. 1 can be predicted for a number of simple cases. First, when the applied force is zero and the initial rate of motion is nonzero, the motion will decay to zero under the influence of friction. Next, when the applied force exceeds the magnitude of the friction, the motion will accelerate in direct proportion to the difference between the applied and friction forces. For these two cases the model is seen to provide a reasonable representation of motion in the presence of friction. Examining next the case where the applied force is less in magnitude than the friction and the initial rate is zero, the model is seen to deviate from the physical law because the modeled friction force is zero for the zero rate condition and an erroneous acceleration of the inertia results.

Assuming the conventional model is evaluated using fixed step size numerical integration, the zero rate case above produces a friction force which, because it exceeds the applied force, causes a rate reversal and leads to a sustained oscillatory process. While the amplitude of the rate excursions can be reduced through a reduction of integration step size, it will be seen that regardless of step size, the rate oscillates about a nonzero mean due to the nonzero input applied force. A special computation is thus necessary to determine a step size sufficiently small to control both the mean and amplitude of rate error. In the 70-m antenna axis servos the ratio of friction to inertia is 102 millidegrees/sec^2 (1.77 milliradians/sec^2) for the azimuth axis and roughly 1.5 times that ratio for elevation. It can be shown that controlling the above rate errors to less than 0.1 mdeg/sec requires a step size of roughly 1.0 msec, which is unreasonably small and leads to excessive computation time and data storage for small computer-based simulations.

II. Derivation of Equations for Modeling Friction

From the foregoing discussion it is evident that accurate computer modeling of motion involving friction is based on knowledge of both the applied force and the velocity of the body influenced by the friction. The velocity determination is thus an essential adjunct of any friction model. When that velocity is determined by a finite step size integration process, the effects of friction reversals resulting from mid-integration-step zero crossings of velocity must be considered. The modeling problem thus becomes the determination of net effective impulse such that the velocity change resulting from the finite step integration is reasonably accurate.

The derivation of the equations for modeling friction is equally applicable to translational or rotational systems. For rotation, the derivation assumes a slowly varying externally applied torque to a constant inertia body in the presence of both static and invariant Coulomb friction and a known, fixed integration interval. The following conditions are considered separately.

(1) The applied torque is greater than the static friction torque and the inertia is initially at rest.

(2) The applied torque and initial rate are such that the rate of motion will not reach zero within the next integration interval.

(3) The applied torque is less than the static friction and the initial rate of motion is such that the rate will reach zero within the integration interval.

(4) The applied torque is greater than the static friction and the initial rate of motion is such that the rate will reach zero within the integration interval.

For Condition (1) above, the net torque acting on the inertia is simply the applied torque diminished in magnitude by the Coulomb friction torque. The effective friction torque, T_f, is a constant with direction opposite to the applied torque

$$T_f = -F_c \cdot \text{sign}\,(T_{ap}) \tag{1}$$

for $\dot{\theta} = 0$ and $|T_{ap}| > F_c$ where F_c is the Coulomb friction torque, the sign function is unit amplitude with the algebraic sign of its argument, T_{ap} is the applied torque, and $\dot{\theta}$ is the rate of

motion. The use of the Coulomb rather than the static value in this case is based on the assumption of an instantaneous transition from the static to the sliding friction case. It will be seen that this assumption results in a minimum net torque equal to the difference between the static and Coulomb values. The resulting rate impulse can be adjusted to better comply with known physical behavior by selection of the integration step size.

The necessary condition for (2) above is determined from the equation of motion in the presence of friction

$$\dot{\theta}(t) = \dot{\theta}(0) + \left(\frac{1}{J}\right)(T_{ap} + T_f)\,t \tag{2}$$

where for rotational motion, J is the inertia moment, $\dot{\theta}(t)$ is the rate at time t, T_{ap} is the applied torque and T_f, the friction torque. Substituting $-F_c \cdot \text{sign}(\dot{\theta})$ for the friction, T_f, solving for $\dot{\theta}(t_{oc}) = 0$, and dividing by the integration step size, t_i, yields

$$\frac{t_{oc}}{t_i} = \frac{-J\,\dot{\theta}(0)}{T_{ap} - F_c \cdot \text{sign}\,[\dot{\theta}(0)]} \tag{3}$$

Negative values of t_{oc}/t_i imply a level of applied torque in excess of the friction and in the same direction as the rate. Positive values imply an applied torque either in a direction opposite the rate, or having a magnitude less than the friction, or both. A negative or unity or greater than unity value of t_{oc}/t_i is a necessary and sufficient condition for Condition (2) above. The net torque in this case is the algebraic difference between the applied and friction torques where the friction is opposite in direction to the rate.

$$T_f = -F_c \cdot \text{sign}\,(\dot{\theta}) \tag{4}$$

for $t_{oc}/t_i < 0$ or $t_{oc}/t_i \geq 1$.

In Condition (3) above, the rate will reach zero at some time within the integration interval and the applied torque will be insufficient to produce a rate in the reversed direction. Because the net torque acts on the inertia for the full interval, it must then decelerate the inertia to precisely zero rate at the end of the interval. The required net torque and necessary conditions are thus

$$T_{\text{net}} = \frac{-J\,\dot{\theta}}{t_i} \tag{5}$$

for $|T_{ap}| < F_s$ and $0 < t_{oc}/t_i < 1$.

In Condition (4) above, the applied torque is sufficient to overcome the static friction level and reverse the rate within the integration interval. The actual friction torque in the physical system will thus reverse coincidental with the rate reversal. The effective friction torque is obtained by averaging the instantaneous friction torque over the integration interval. Thus

$$T_f = -F_c \cdot \text{sign}(\dot{\theta}) \left(\frac{[t_{oc} - (t_i - t_{oc})]}{t_i}\right)$$

$$T_f = F_c \cdot \text{sign}(\dot{\theta}) \left[1 - 2\left(\frac{t_{oc}}{t_i}\right)\right] \tag{6}$$

for $|T_{ap}| > F_s$ and $0 < t_{oc}/t_i < 1$.

If the initial rate is zero and Condition (1) above is not satisfied, the friction equals the applied torque and the net torque becomes zero. Further, since Conditions (1) through (4) encompass all possible torque and rate conditions of interest, Eqs. (1) through (6) together with their conditions of applicability form the basis for defining effective friction torque and the net torque.

III. Application to Practice

A function block incorporating the logic and equalities of Eqs. (1) through (6) was developed for incorporation into a dynamic simulation model of the 70-m azimuth axis servo using MATRIXx, a copyrighted software program from Integrated Systems, Inc. for simulation of dynamic systems. The friction model utilizes four general equation building blocks and one standard function block from the MATRIXx utilities. Because the simulation program does not facilitate conditional branching in function blocks, it was necessary to structure the algorithm to employ eight logical variables whose one/zero values define Conditions (1) through (4) discussed in relation to Eqs. (1) through (6). The logical variables are then used in one equation for the net torque.

The algorithm inputs are $U1$, applied torque, $U2$, output rate from an adjoint integration process, and $U3$, a unit variable with the algebraic sign of $U2$. The output is $Y22$, the net torque to the integrator. Parameters are the static and Coulomb friction levels, F_s and F_c, the inertia moment, J, and the integration interval t_i.

The computations are grouped in function blocks as shown in Fig. 2 to avoid intermixing relational and arithmetic opera-

tors. The logical assignments use the convention where the lefthand variable is true (one) if the righthand condition is satisified, and false (zero) otherwise. A listing of variables and equations is provided below.

Friction logical variables

$$YY1 = U1 > F_c$$

$$YY2 = U1 < -F_c$$

$$YY3 = U2 > 0$$

$$YY4 = U2 < 0$$

$$YY5 = U1 > F_s \text{ and not } (YY3 \text{ or } YY4)$$

$$YY6 = U1 < -F_s \text{ and not } (YY3 \text{ or } YY4)$$

$$YY7 = YY1 \text{ or } YY2$$

Numeric variables, $YN1$ = critical torque, $YN2 = t_{oc}/t_i$

$$YN1 = U2 \frac{J}{t_i}$$

$$YN2 = \frac{YN1}{U3F_c - U1}$$

Logical variable

$$YY10 = YN2 > 0 \text{ and } YN2 < 1$$

Algebraic friction equation

$$Y21 = (2 \cdot YN2 - 1) F_c$$

$$\begin{aligned} Y22 = {} & [YY3 (U1 - Y21) + YY4 (U1 + Y21)] \; YY7 \cdot YY10 \\ & + [YY3 (U1 - F_c) + YY4 (U1 + F_c)] \; (1 - YY10) \\ & + YY5 (U1 - F_s) + YY6 (U1 + F_s) \\ & - (1 - YY7) YY10 \cdot YN1 \end{aligned}$$

IV. Simulation Test Results

Performance of the conventional friction model of Fig. 1 and the new model of Fig. 2 was compared in a dynamic simulation of the 70-m azimuth axis position servo. To simplify the simulation results, the flexible dynamics of the antenna structure were replaced with equivalent rigid-body parameters, thereby reducing the dynamic system to eighth order. The simulations are otherwise representative of actual system performance. The system excitation was a small (1.0 millidegree) position step transient. The control torque, measured in units of psi of hydraulic differential pressure, and axis rate in millidegrees/sec were recorded for comparison. The simulation was run for a total time of 5.0 sec with a 10 msec integration step size for both friction models. Results for the conventional model are shown in Fig. 3 and for the new model in Fig. 4.

The position loop dynamics simulated are such that the control torque changes slowly in response to the small transient applied here. The oscillatory behavior of the conventional friction model is evident during the intervals when the applied torque is less than the friction. The nonzero mean rate during these intervals is erroneous as the rate should be zero until the applied torque exceeds the 400 psi friction threshold. The irregularities on the rising and descending portions of the torque graph appear to be the spurious result of the oscillations coupling back through the rate loop.

The new friction model produces smooth torque transitions and zero rate in the intervals between the static friction levels (425 psi) in conformance with expectations based on the physical laws of friction. The ripple in the rate result is most likely the 7.0-Hz mode of the gear actuator stiffness included in the model.

V. Summary and Conclusions

An improved method for modeling dynamic motion in the presence of friction has been described. Simulation test results demonstrated that the anomalies of more conventional methods are corrected without increasing computer processing time. While the new algorithm is based on an external Euler integration, it should be capable of extension to incorporate a trapezoid- or possibly a polynomial-based integration method. The increased complexity of the predictive calculation with a polynomial integration may, however, negate any advantage to be gained with the more efficient integration methods.

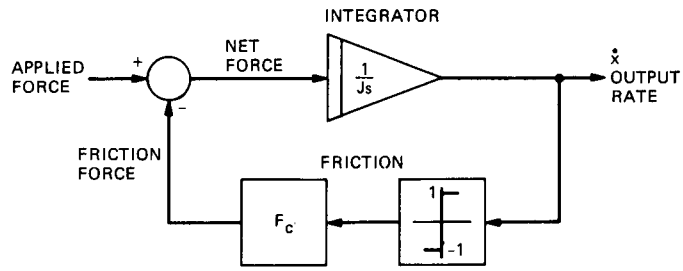


Fig. 1. Conventional friction model for single coordinate motion.

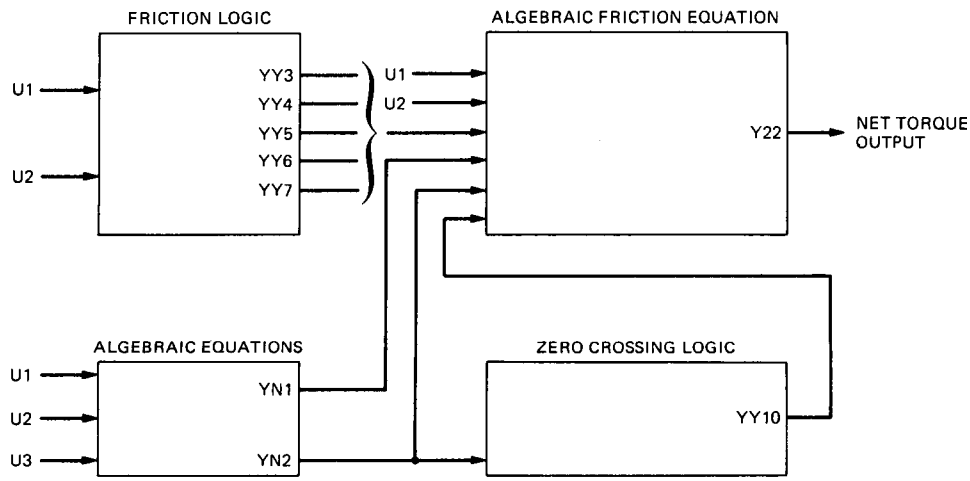


Fig. 2. Simulation function block implementation of the new algorithm.

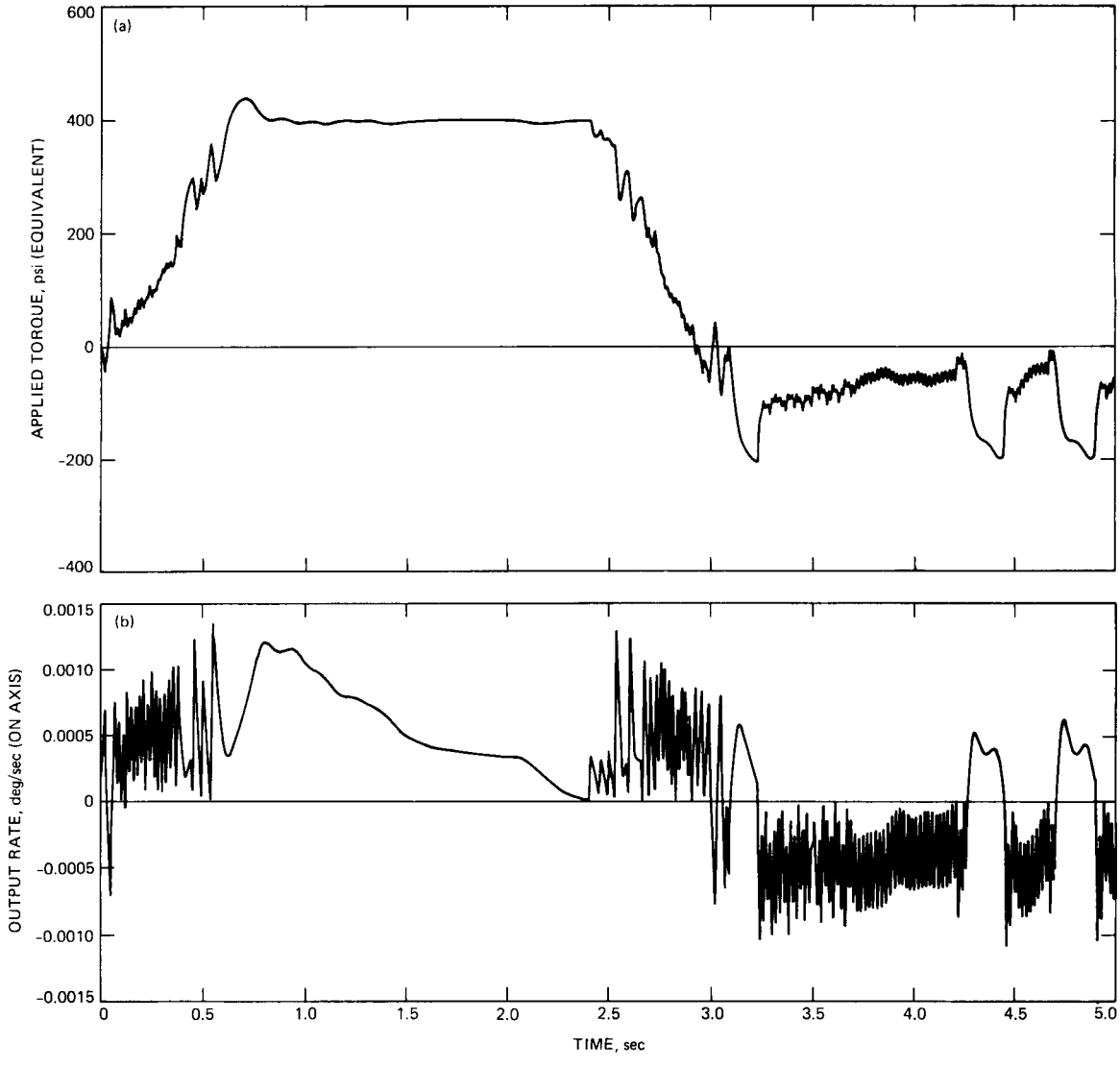


Fig. 3. Conventional friction simulation algorithm: (a) applied torque and (b) output rate.

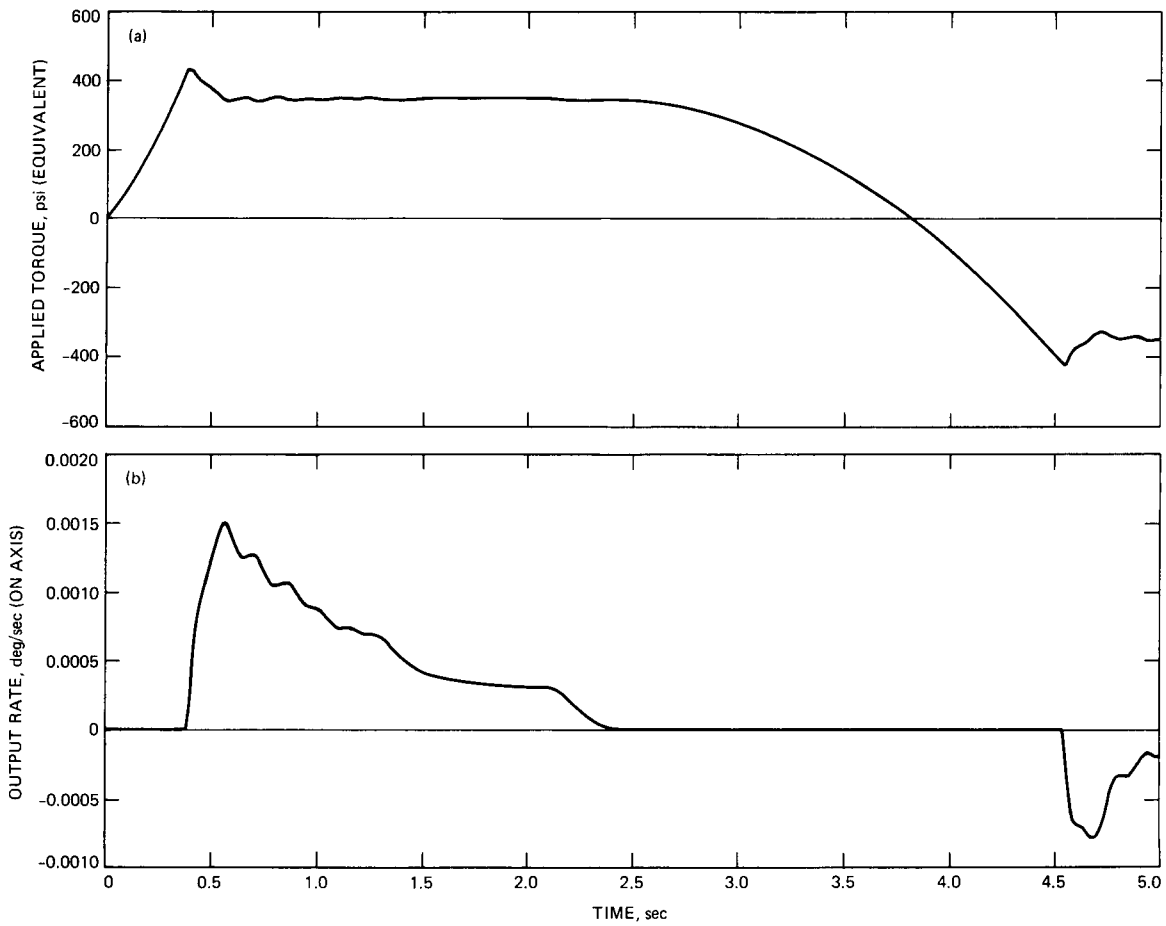


Fig. 4. New friction simulation algorithm: (a) applied torque and (b) output rate.

Theoretical Comparison of Maser Materials for a 32-GHz Maser Amplifier

J. R. Lyons
Radio Frequency and Microwave Subsystems Section

This report presents the computational results of a comparison of maser materials for a 32-GHz maser amplifier. The search for a better maser material is prompted by the relatively large amount of pump power required to sustain a population inversion in ruby at frequencies on the order of 30 GHz and above. The general requirements of a maser material and the specific problems with ruby are outlined. The spin Hamiltonian is used to calculate energy levels and transition probabilities for ruby and twelve other materials. A table is compiled of several attractive operating points for each of the materials analyzed. All the materials analyzed possess operating points that could be superior to ruby. To complete the evaluation of the materials, measurements of inversion ratio and pump power requirements must be made in the future.

I. Introduction

This report describes the results of a theoretical evaluation of several paramagnetic materials being considered for use in a 32-GHz maser amplifier. Previously, ruby has been very successfully employed in 2.3- and 8.4-GHz masers [1], and in an 18- to 26-GHz tunable maser [2]. However, due to a monotonically decreasing inversion ratio above 12 GHz for ruby (for push-pull pumping) ruby becomes less favorable at higher frequencies. The inversion ratio is defined as the ratio of the inverted-spin population difference with the pump on to the thermal-equilibrium population difference with the pump off, and is determined experimentally by the ratio of gain (dB) with the pump on, to absorption (dB) with the pump off. Moore and Neff [3] and Shell (private communication) measured an inversion ratio of 1.1 at 32 GHz. In a recently completed reflected-wave maser [5], the inversion ratio was estimated to be 0.7 to 0.8. This low inversion ratio was a contributing factor to the reduced gain-bandwidth of the maser compared to a similar 22-GHz maser [31] in which the inversion ratio was at least 1.6. Theoretical calculations of spin-lattice relaxation rates [29] suggest that the low inversion ratio ($I \approx 1$) is inherent in ruby at this operating point ('operating point' refers to a given dc field strength, crystal orientation, and pumping scheme). Hence, other paramagnetic materials, as well as other operating points of ruby, are investigated as a first step toward finding the best maser material at 32 GHz.

Before proceeding with the materials evaluation, two points should be made. First, the inversion ratio in the 32-GHz reflected-wave maser (RWM) could probably be improved by either using more pump power or by using a maser structure with a higher Q at the pump frequency. In the former case, heating of the maser due to microwave losses will degrade the

gain and noise temperature performance, but to what extent is unknown. In the latter case, the resonant structure would significantly reduce the tunability of the RWM, but would improve the pump power coupling over a still useful instantaneous bandwidth. However, even with these engineering improvements, the inversion ratio would still only approach unity.

The second point is that the inversion ratio, even though central to the evaluation of a maser material, cannot be accurately addressed by the methods presented here. It is believed that the low inversion ratio of ruby at the 32-GHz RWM operating point is a result of an unfavorable set of spin-lattice relaxation rates. These relaxation rates and the corresponding inversion ratios have been calculated, but, due to the complexity of the calculations, will be presented in a future report.

In Section II the requirements of a maser material are discussed, including the difficulties with ruby. A list is provided of the materials evaluated. In Section III the use of the spin Hamiltonian to calculate energy levels and transition probabilities is outlined. In addition, limits on the inversion ratio and a material figure-of-merit are discussed and the method of computation is reviewed. In Section IV a table of several promising operating points is presented for each of the materials analyzed. The conclusions are presented in Section V. The Appendix contains a table of measured relaxation times obtained from the literature for several materials of interest.

II. Maser Material Considerations

The choice of a maser material is the single most important factor in maser design. The maser material consists of a nonmagnetic crystalline lattice lightly doped (0.01–0.1 percent) with paramagnetic ions. Detailed discussions of suitable paramagnetic ions and host crystals are given elsewhere [6], [7]. Here, only an outline of the necessary and desirable material properties is given.

The most common paramagnetic ions are transition metals and rare earths, because of their unfilled $3d$ and $4f$ electron shells. To operate in CW mode, the ion should possess an orbital ground state with three or more spin levels. This eliminates most of the rare earths. The additional requirement of a negligible nuclear magnetic moment (a source of inhomogenous broadening) reduces the possible ions to Cr^{3+}, Fe^{3+}, Ni^{2+}, and Gd^{3+}. (Actually, it is not clear that such broadening would adversely affect a maser with an inhomogenous applied field.) Only Cr and Fe are considered here, as they are by far the most common choices of active ions. The electronic configurations of Cr^{3+} and Fe^{3+} are $3d^3$ and $3d^5$, respectively. The corresponding free-ion ground states are $^4F_{3/2}$ and $^6S_{3/2}$. Hence, Cr^{3+} has four spin levels and Fe^{3+} has six.

The host crystal must be non-magnetic, non-metallic, and available in large single crystals of a high degree of perfection. The material should have a sufficiently high thermal conductivity and small loss tangent at liquid helium temperatures and microwave frequencies to minimize heating of the lattice. To facilitate the microwave engineering, the crystal should possess a relatively isotropic and temperature-independent dielectric constant. Finally, the material should be machinable and chemically stable, and be able to withstand thermal cycling between liquid helium and room temperatures.

The active ion substitutes for one of the metal ions in the crystal lattice. The local crystal electric field seen by the ion splits the highly degenerate orbital ground state into degenerate pairs (assuming the number of spin levels is even). This splitting of spin levels due to the crystalline electric field is termed the zero-field splitting (ZFS). The ZFS must be large enough to permit pump-induced transitions between nonadjacent spin levels. As a rule-of-thumb, the ZFS should be of the same order of magnitude as the signal frequency.

Another very important material parameter is the spin-lattice relaxation time, which describes how long spins remain in an excited state before returning to thermal equilibrium with the lattice. At microwave frequencies, spin-lattice interaction is the dominant spin relaxation mechanism. For relatively low concentrations of paramagnetic ions (<0.05 percent) and at liquid helium temperatures, the most significant spin-lattice interaction is thought to be the Kronig-Van Vleck mechanism [8], in which lattice vibrations induce transitions between spin states, and spin-spin interactions are neglected. The spin-lattice relaxation times of the various transitions must be long enough ($\geqslant$ msec) to permit saturation of the pumped levels with a reasonable amount of pump power. Impurities, ion clustering, and dislocations (all of which are a function of the crystal growth procedure) can shorten relaxation times, so pure, defect-free crystals are preferable.

Another material parameter is the (unbroadened) linewidth Δf_L of the material. At liquid helium temperatures, Δf_L is determined primarily by spin-spin interactions and is inversely proportional to the spin-spin relaxation time, which is the average length of time between random dephasing "collisions" of neighboring spins [9]. The linewidth is usually within the range of 10–100 MHz for solid-state maser materials. For linear stagger-tuned masers of bandwidth $\gg \Delta f_L$, Δf_L can be shown to have no first-order influence on the gain-bandwidth properties of the maser (see Section III). However, if the taper is along the length of the material, a material with a smaller Δf_L may exhibit a larger noise temperature at one end of its band-

pass [9]. At present, it is not understood how Δf_L impacts pump power requirements.

Finally, many materials possess two or more magnetically inequivalent sites (i.e., sites having different spectra) for the active ions to occupy, thus decreasing the density of useful ions. (Gain in dB is proportional to spin density.) In most materials, certain orientations exist for which these sites become equivalent. For such materials, only these orientations will be analyzed.

As mentioned in Section I, difficulties were experienced using ruby at 32 GHz. A relatively low inversion ratio was obtained for the orientation employed in the 32-GHz RWM. The pump transitions in ruby at this orientation are quite weak, making it difficult to saturate the levels. This is a result of the small zero-field splitting (ZFS) of ruby: ZFS = 11.4 GHz, which is only about one-third of the signal frequency. To obtain sufficient separation between spin levels for amplification at 32 GHz, a relatively large magnetic field (11.8 kG) must be applied. This field becomes the dominant influence on the spins, far exceeding the effects of the local crystal field. Under such conditions, the spins in the lattice assume nearly pure-spin characteristics, as if the Cr ions existed freely in the magnetic field. The selection rules of quantum mechanics allow transitions only between adjacent pure-spin states [10], thus leading to a small stimulated transition probability for the pump transitions in ruby. A material with a larger ZFS will in general have stronger pump transitions and will therefore better absorb the pump power, all other factors remaining the same.

The materials analyzed in this work are listed below. Details of the crystal structure, orientation of magnetic axes, spin Hamiltonian, and site equivalence are given in the references.

Ruby (Al_2O_3:Cr)	[11]
Emerald ($Be_3Al_2Si_6O_{18}$:Cr)	[12]
Spinel:Cr ($MgAl_2O_4$:Cr)	[13]
YAG:Cr ($Y_2Al_5O_{12}$:Cr)	[14]
YGG:Cr ($Y_2Ga_5O_{12}$:Cr)	[14]
Rutile:Cr (TiO_2:Cr)	[15]
Zinc Tungstate:Cr ($ZnWO_4$:Cr)	[16]
Andalusite:Cr (Al_2SiO_5:Cr)	[17]
Yttrium Oxide:Cr (Y_2O_3:Cr)	[18]
Rutile:Fe (TiO_2:Fe)	[19]
Zinc Tungstate:Fe ($ZnWO_4$:Fe)	[20]
Andalusite:Fe (Al_2SiO_5:Fe)	[21]
Sapphire:Fe (Al_2O_3:Fe)	[22]

III. Evaluation of the Spin Hamiltonian and Associated Parameters

The spin Hamiltonian H_s describes the interaction of the electron spin of the paramagnetic ion with the local crystal field and with the applied magnetic field. Evaluation of H_s allows calculation of the energy levels and transition probabilities for the spin system of a singlet orbital ground state ion in a radiation field. Detailed discussions of the spin Hamiltonians may be found in the literature [8], [23]. Berwin [24] gives a detailed derivation of the spin Hamiltonian based on the formulation of Bleaney and Stevens [23] for ruby. In [25], Berwin discusses the spin Hamiltonians for several of the materials listed in Section II.

For most materials of interest, the assumption of an "intermediate" crystal field is made, meaning that the interaction energy of the crystal field with the ion falls between the Coulomb and spin-orbit interaction terms. In deriving H_s, the spin-orbit and Zeeman terms are treated together as a perturbation on the singlet orbital ground state. The absence of orbital degeneracy is sufficient for quenching of the orbital angular momentum L [8]. That is, to first-order, L is equal to zero, so the ion behaves in a pure-spin-like manner. Up to a second-order perturbation, the spin-orbit coupling admixes the singlet ground state with higher-lying orbitals, restoring some of the orbital angular momentum. This second-order effect is the source of the ZFS.

For ruby, the spin Hamiltonian has the form [11]

$$H_s = g\beta\bar{B}\cdot\bar{S} + DS_z^2 \tag{1}$$

where the spectroscopic splitting-factor is $g \approx 2$ ($g = 2.0023$ for pure spin), β is the Bohr magneton, $\bar{B}$ is the applied dc magnetic field, and

$$\bar{S} = \hat{x}S_x + \hat{y}S_y + \hat{z}S_z \tag{2}$$

is a vector of spin operators. The spin operators describe the observable properties of the paramagnetic ion spin states and may be conveniently written in matrix form [10]. The Cartesian directions are along the principal axes of the magnetic complexes. These axes, usually expressed in terms of the crystallographic axes, are used to describe the symmetry of the magnetic resonance spectrum. The orientation of $\bar{B}$ is expressed by the usual azimuthal and polar angles, ϕ and θ. The constant D determines the ZFS and reflects, in principle, the

extent of admixing with higher lying orbitals. This spin Hamiltonian exhibits axial symmetry about the magnetic z-axis and has a ZFS = $2|D|$.

For magnetic complexes of lower symmetry, additional spin operator terms may be needed to accurately specify the resonance spectrum; e.g., $(S_x^2 - S_y^2)$ and $(S_x^4 + S_y^4 + S_z^4)$, the so-called orthorhombic and cubic terms [26]. The form of the required spin operator terms can sometimes be determined from crystal field theory through the use of equivalent operators, as discussed in [24]. However, the crystal field approach usually assumes ionic bonding and neglects covalency effects. The fact that the coefficients of H_s, and quite often the form of H_s, must be determined experimentally (by fitting to EPR data) is quite likely a result of this assumption.

Given an H_s, one can solve for the spin energy levels and eigenstates by solving:

$$H_s|\psi_i\rangle = E_i|\psi_i\rangle \qquad i = 1, 2, \ldots, 2S+1 \tag{3}$$

where $|\psi_i\rangle$ is usually written as a linear combination of pure-spin states,

$$|\psi_i\rangle = a_i|S\rangle + b_i|S-1\rangle + \ldots + r_i|-S\rangle \tag{4}$$

and where E_i is the energy of the ith level, and S is the spin of the ion. The labeling of the pure-spin states is identical to the labeling of states for a large applied field. Equation (3) is most easily solved by expressing it in matrix form and using the usual matrix methods to solve for the eigenvalues and eigenvectors of H_s. Note that H_s is Hermitian.

Knowing the eigenstates, one may then calculate the rate of stimulated transitions due to an RF magnetic field $\bar{H}_1$. Applying Fermi's golden rule, the probability of a transition between states i and j is [9]:

$$W_{ij} = \frac{1}{4}\gamma^2 g(f)\,|\langle\psi_i|g\beta\bar{S}\cdot\bar{H}_1|\psi_j\rangle|^2 \tag{5}$$

where $\gamma = g\beta\mu_0/\hbar$, and $g(f)$ is the line shape (as a function of frequency) for the transition. The term sandwiched in the matrix element is the magnetic dipole interaction energy. Since non-spin operators may be removed from the matrix element in Eq. (5), it is convenient to define the vector quantity

$$\bar{\sigma}_{ij} = \langle\psi_i|\bar{S}|\psi_j\rangle \tag{6}$$

The vector components are defined by the relation:

$$\bar{\sigma}_{ij} = \frac{1}{2}(\alpha_{ij}\hat{x} + \beta_{ij}\hat{y} + \gamma_{ij}\hat{z}) \tag{7}$$

where α, β, and γ are in general complex numbers. Noting Eq. (4), α, β, and γ are easily evaluated for any number of spin levels; in [24] the results for a 4-level spin system are given.

As a means of comparing the relative strength of transitions, Siegman [9] has defined the quantity

$$\sigma^2 = \frac{\bar{H}_1^* \cdot \bar{\sigma}\bar{\sigma}^* \cdot \bar{H}_1}{|H_1|^2} \tag{8}$$

The stimulated transition probability can then be written as:

$$W_{ij} = \frac{1}{4}\gamma^2 g(f)\,|H_1|^2\,\sigma_{ij}^2 \tag{9}$$

The absolute maximum value of σ^2 is obtained by using RF fields that are polarized such that $\bar{H}_1$ is paralled to $\bar{\sigma}$. This value of σ^2 is given by the trace of $\bar{\sigma}\bar{\sigma}^*$ [9], which has the value:

$$(\sigma^2)_{max} = \frac{1}{4}(|\alpha|^2 + |\beta|^2 + |\gamma|^2) \tag{10}$$

This expression will be used to compare transition strengths. This formulation does not consider whether or not the prescribed polarization of $\bar{H}_1$ is achievable in a given microwave circuit.

The gain of the maser is obviously of central importance and involves several important material parameters. The gain in dB of an unbroadened RWM or TWM [9] is

$$G_{dB} = 27.3\,\frac{s\ell}{Q_m} \tag{11}$$

where s is the slowing factor, ℓ is the maser length in free-space wavelengths, and Q_m is the magnetic Q of the maser material. The magnetic Q is defined as

$$Q_m = \frac{\text{energy stored in material}}{\text{energy emitted per cycle by material}} \tag{12}$$

Assuming $\frac{hf}{kT} \ll 1$ ($\frac{hf}{kT} \approx 1/3$ for $f = 32$ GHz, $T = 4.2$ K), the reciprocal of Q_m may be expressed as [9]

$$\frac{1}{Q_m} \approx \frac{\gamma^2\hbar}{\pi\mu_0}\cdot\frac{hf}{kT}\cdot\frac{I\sigma^2\eta}{\Delta f_L}\cdot\frac{N}{\text{no. of levels}} \tag{13}$$

where I is the inversion ratio, η is the filling factor, and N is the spin density. The filling factor accounts for the fraction of RF field in the material and the degree to which the field is optimally polarized; its value varies between 0 and 1. The spin density is determined by the concentration of paramagnetic ions; for 0.05 percent Cr concentration in ruby, $N = 2.35 \times 10^{19}$ spins/cm^3. From Eqs. (11) and (13), we may express G_{dB} in terms of material parameters as:

$$G_{dB} \propto \frac{I\sigma^2 N}{\Delta f_L \cdot \text{no. of levels}} \tag{14}$$

The inversion ratio is defined as

$$I = \frac{\Delta n_{ij}}{\Delta N_{ji}} \tag{15}$$

where $\Delta N_{ji} = N_j - N_i$ is the thermal-equilibrium population density-difference, and Δn_{ij} is the population-density difference under pumped conditions. To determine I for the signal transition of a multi-level spin system, one must solve a set of rate equations that accounts for both stimulated transitions and spin relaxation [9]. These rate equations may be simplified by assuming steady-state conditions and saturated pump levels, and by neglecting the influence of the incoming signal. Since the relaxation rates are not known (this issue will be addressed in the aforementioned future report), the rate equations cannot be solved for the actual inversion ratio. Instead, assumptions are made about the relaxation rates and the corresponding I is determined. In one case, all relaxation rates are assumed to be equal and the inversion ratio is defined as I_{equal}. An upper limit can be put on I by assuming an optimum set of relaxation rates exist [9]. This is defined as I_{opt}. Note that the actual inversion ratio may be less than both I_{opt} and I_{equal}, but for pumping schemes employing two pumps, the actual inversion ratio often lies between I_{opt} and I_{equal}. Expressions for I_{opt} and I_{equal} are easily derived from the rate equations but will not be given here.

A maser material "figure-of-merit," indicating the gain-bandwidth potential of the material, was proposed [27] to be $\Delta f_L/Q_m$. From Eq. (13), and considering only material parameters, we find that

$$\frac{\Delta f_L}{Q_m} \propto \frac{I\sigma^2 N}{\text{no. of levels}} \tag{16}$$

Hence, the figure-of-merit optimizes the product $G_{dB} \cdot \Delta f_L$. For the case of a linear stagger-tuned maser with bandwidth $\Delta f \gg \Delta f_L$, one can show that the quantity on the right-hand side of Eq. (16) optimizes the product $G_{dB} \cdot \Delta f$.

The computer analysis was performed on the JPL UNIVAC (F-system) with a modified version of an existing Fortran code [25] originally written for the analysis of ruby. The JPL Fortran subprogram HERMQR[1] was used to compute the eigenvalues and eigenvectors of H_s. The existing code was modified to include the spin Hamiltonians of the other materials. Other small modifications were also made.

The inputs to the program are as follows:

(1) material

(2) range of θ and θ-increment

(3) a single value of ϕ

(4) range of B and B-increment

(5) signal frequency window f_{LO}, f_{HI}

(6) minimum allowed value for max (α, β, γ) for signal transition, σ_{min}

The code is run for a given material at a given ϕ. Typical values are θ = 0 to 90 deg, $\Delta\theta$ = 10 deg, 0 deg $\leqslant \phi \leqslant$ 90 deg, and B = 0 to 15 kG, ΔB = 0.5 kG. For a 32-GHz signal frequency, the window was usually f_{LO} = 31 GHz and f_{HI} = 33 GHz. For transitions falling within this range, σ_{min} = 1.0 was chosen. The signal frequency window and σ_{min} are used to pre-select operating points, thus decreasing computer output.

The program outputs are the following:

(1) energy levels and eigenstates

(2) transition probabilities (both α, β, γ and σ^2)

(3) local values for $\frac{\Delta f_s}{\Delta B}$ and $\frac{\Delta f_p}{\Delta f_s}$

(4) I_{opt}, I_{equal}, and figure-of-merit

Because of the large number of possible pumping schemes (especially for the 6-level systems), the following guidelines were employed in choosing schemes:

(1) signal transition is between adjacent levels

(2) use two pumps, when possible

(3) pumps can skip one level at most

These guidelines limit the pumping schemes to the usual ones employed. Many other schemes are possible [9].

[1] *JPL Fortran V Subprogram Directory*, Fifth Edition, JPL Publication D-829 (internal document), Jet Propulsion Laboratory, Pasadena, California, July 1982.

IV. Results and Discussion

Table 1 compares several of the more promising operating points of each of the materials analyzed, starting with ruby. The materials are arranged in order of increasing ZFS; for materials with $S = 5/2$, the ZFS of the lower degenerate states is used. Across the top of the table is the material name and operating point number. The first row of the table shows the paramagnetic ion used. The second row gives the ZFS. Note that all the materials have a ZFS larger than ruby. The third row indicates the number of magnetically non-equivalent ionic sites in the lattice; a "1" means that all sites are equivalent. The fourth row gives the orientation of B in terms of the polar and azimuthal angles, θ and ϕ, measured with respect to the axes of the magnetic complex. If a value for ϕ is not given, then the Hamiltonian is axially symmetric. For materials with non-equivalent sites, θ and ϕ are restricted to values for which the sites are equivalent. The fifth row of the table gives the magnitude of B. The fields do not exceed 14 kG for the operating points shown.

Rows 6, 7, and 8 list the signal and pump frequencies with the corresponding transition levels shown in parentheses. The signal frequency is always 32.0 GHz. The pump frequencies vary roughly between 50 and 90 GHz. Larger pump frequencies will in general yield larger values of I_{opt} and I_{equal}. On the other hand, copper and dielectric losses increase at higher frequencies, with the result that heating of the maser structure may restrict the use of high pump power levels at high frequencies for some materials. Note also that pump frequencies within the same waveguide band simplify engineering issues.

Row 9 gives the value of $\Delta f_s/\Delta B$ evaluated near the operating point. For a maser tunable over a wide range, $\Delta f_s/\Delta B$ should be of the same sign and of similar magnitude for the maser material and the isolator material. Rows 10 and 11 give values of $\Delta f_p/\Delta f_s$ (actually, $\Delta f_p/\Delta B \cdot \Delta B/\Delta f_s$) for both pumps. This parameter is indicative of the pump bandwidth required for a given signal bandwidth, so it is preferable for $|\Delta f_p/\Delta f_s|$ to be as small as possible. For most of the operating points $\Delta f_p/\Delta f_s \approx 2$, but several have values <1. Note that $\Delta f/\Delta B$ is evaluated as a simple two-point difference, with the second point arbitrarily located 200 G from the operating point. For strongly curved energy levels, these values may not be accurate across the desired band.

Rows 12, 13, and 14 give σ^2 for the signal and pump transitions for optimum elliptically polarized fields according to Eq. (10). Recall from Eqs. (14) and (16) that G_{dB} and the figure-of-merit are proportional to σ_s^2, so as large a value of σ_s^2 as possible is desired. In general, σ_s^2 is a factor of 2- to 3-times larger for the 6-level spin systems. Similarly, a large value of σ_p^2 is preferred, since the pump power required for saturation is inversely proportional to σ_p^2. According to [29], the pump power required for saturation will satisfy

$$P_{pump} \propto \frac{f_p \cdot \Delta f_p}{\sigma_p^2 \tau_p} \tag{17}$$

where τ_p is the effective pump relaxation time; τ_p is not identical to the measured pump relaxation time. The values of σ_p^2 in Table 1 span nearly two orders of magnitude.

Rows 15 and 16 show the inversion ratios for equal and optimum relaxation times. The values of I_{equal} and I_{opt} are similar for the various operating points, except when only one pump is employed.

Finally, row 17 gives the material's figure-of-merit, computed in units of MHz, as

$$\frac{\Delta f_L}{Q_m} = \frac{5.6\, I_{opt} \cdot \sigma_s^2}{\text{no. of levels}} \tag{18}$$

This follows from Eq. (13) evaluated at $f = 32$ GHz, $T = 4.2$ K, $\eta = 0.5$, and $N = 2.35 \times 10^{19}$ spins/cm^3. Since the true inversion ratio for a given operating point may be as much as a factor of 3 or more smaller than I_{opt}, a detailed comparison of figure-of-merits could be misleading.

Table 1 is by no means complete in the sense that one may confidently select the best maser material from it. Two very important parameters are missing: the actual inversion ratio and the pump power, P_{pump}, required to maintain that inversion ratio. At present, both of these parameters must be measured.

Since both I and P_{pump} depend critically on relaxation times, a table of relaxation times, located in the Appendix and labeled Table A, was compiled from data found in the literature. Ionic concentration, frequency, orientation, and transition information is included. Because these parameters do not coincide with pump operating-points of interest to us, and because of the dependence of relaxation times on measurement technique and crystal growth procedures [30], the data in Table A could easily be an order of magnitude or more different from what would be measured for the materials in Table 1. Hence, the relaxation times in Table A are not used in any calculations in this work, even though they are the best values available to us at the present time.

Before discussing the many materials in Table 1, consider the operating point in which ruby is presently being used at 32 GHz (first column of table, Ruby No. 1). The ruby is oriented at the double-pump angle ($\theta = 54.7$ deg) and pumped

in the push-pull mode, so the pump frequencies are equal. Scanning down the column, two potential problems can be seen with this operating point. First, the pump bandwidths are nearly twice the signal bandwidths, so if 500 MHz of signal bandwidth is desired at 32 GHz, 1-GHz bandwidth must be pumped at 66 GHz. To pump such a large bandwidth, the pumps must be swept across the band, effectively reducing the pump power at a given frequency. How detrimental this is depends on the relaxation times of the pump transitions.

The second problem with ruby at this operating point is the weak pump transitions: $\sigma_{p1}^2 = 0.05$ and $\sigma_{p2}^2 = 0.04$ compared to $\sigma_s^2 = 1.92$. For this reason, high levels of pump power are used in the 32-GHz RWM, although the pumped levels are still not saturated. Note that a small σ_p^2 does not preclude good maser performance, as demonstrated by the 18- to 26-GHz maser of Moore and Clauss [2], [31] for which $\sigma_{p1}^2 = 0.07$ and $\sigma_{p2}^2 = 0.11$.

The most significant problem with ruby, the low inversion ratio, is not indicated by the table. Measured values of I for the case of saturated pump transitions have been approximately 1.1 ([3], and J. Shell, private communication). (The similarity to $I_{equal} = 1.1$ does not necessarily mean that the relaxation times are equal.) In the 18- to 26-GHz range, measured values of I have been in the range of 1.6 to 1.8 [2], [3], [31].

Finally, from Table A it can be seen that ruby has long relaxation times compared to the other materials. Hence, even though σ_p^2 is small, the denominator of Eq. (17) remains large enough for ruby to require large but manageable pump power.

Consider several other operating points in Table 1. Since ruby has worked so well in the past, ruby at another orientation is an obvious candidate for a maser material. The second column of the table, Ruby No. 2, shows ruby at $\theta = 90$ deg and with a push-push pumping scheme. Even though σ_s^2 and σ_{p1}^2 are weaker and the values for I_{equal} and I_{opt} are less than for Ruby No. 1, if the actual inversion ratio is >1.5, Ruby No. 2 could yield a higher gain-bandwidth product. Some investigators [4] claim Ruby No. 2 to be superior to Ruby No. 1 at millimeter wavelengths because of a higher inversion ratio and less critical orientation (less spreading of pump power due to c-axis wander).

The sapphire host has many desirable properties, so Fe-doped sapphire is a logical choice. For Sapphire No. 1, σ_p^2 is 2- to 5-times stronger than that of Ruby No. 1 and σ_s^2 is 2- to 3-times that of Ruby No. 1. However, according to Table A, the pump relaxation times may be an order of magnitude shorter, implying that Sapphire No. 1 could require several times the pump power of Ruby No. 1. Other investigators [28] suggest that the relaxation times of Fe-doped sapphire are similar to those of ruby. If this is true, then Sapphire No. 1 could require several times less pump power than Ruby No. 1. Measurements of the relaxation times and pump power required must be made to determine which scenario is correct.

Emerald has some similarity to ruby, having the same spin Hamiltonian and potentially long relaxation times. If the inversion ratio for Emerald No. 1 is $\geqslant 2$, then this operating point would be very attractive. A problem with emerald is the difficulty of its growth, which may not allow the high degree of crystal perfection necessary.

Zinc tungstate has a complicated H_s, large ZFS values, and may have short relaxation times, making it quite different from ruby. Cr-doped zinc tungstate has several promising operating points. In particular, $ZnWO_4$ No. 1 is attractive, assuming $I \approx I_{opt}$. Fe-doped zinc tungstate exhibits a large number of excellent operating points. For $ZnWO_4$ No. 3, σ_s^2 is 2- to 3-times greater than that of Ruby No. 1 and σ_p^2 is 100 times that of Ruby No. 1. This large value of σ_p^2 raises the question of whether it is preferable to have large σ_p^2 and small τ_p ($ZnWO_4$ No. 3) or small σ_p^2 and large τ_p (Ruby No. 1). Assuming the product $\sigma_p^2 \cdot \tau_p$ is constant, the pump power requirements will be similar, but in the former case more energy would be transferred to the lattice. This could raise the temperature of the maser material, thereby decreasing the gain; however, this possibility has not been considered in detail. Harmonic cross relaxation may be a problem for several of the better operating points for $ZnWO_4$:Fe. One similarity zinc tungstate has with ruby is that it can be grown by the Czochralski method.

The rutiles appear promising, but the large, anisotropic, temperature-dependent dielectric constant of rutile makes it unattractive from an engineering standpoint.

One can easily see from Table 1 that many of the other materials analyzed may make excellent maser materials, but the lack of information on inversion ratios, pump power requirements, relaxation times, etc., makes them difficult to evaluate.

Another possibility, not addressed in Table 1, is to use standard ruby doped with a fast-relaxing impurity. This additional impurity may be added in the melt or created in the finished ruby by exposure to X-rays (so-called orange-ruby [9]). A properly chosen impurity can shorten certain relaxation times, which, by making the times more optimal, can increase the inversion ratio. However, the impurity would not alter the ZFS, so the pump transitions would still be weak.

We hope to eventually make measurements of inversion ratios, pump power requirements, and relaxation times at

32 GHz and around 60 GHz on several of the materials in Table 1.

Other materials we would like to analyze but for which we do not have the spin Hamiltonians are spinel:Fe and chrysoberyl:Cr, Fe.

V. Conclusions and Future Work

Any of the materials analyzed in this work may yield better maser performance than does ruby at 32 GHz at the double-pump angle. However, several key parameters related to pump power requirements may eliminate some or all of these materials. Based on results from the analysis of the spin Hamiltonians and on scanty (and unreliable) relaxation time data, several materials show particular promise (e.g., Fe-doped zinc tungstate).

To complete the materials evaluation, it will be necessary to measure the inversion ratio and pump power required for saturation for each operating-point of interest. Barring cross-relaxation and other concentration-dependent effects, knowledge of the relaxation-times would be sufficient to calculate both I and P_{pump}. However, the subtlety of measuring relaxation-times will most likely require that I and P_{pump} be measured.

Better understanding of the low inversion ratio of ruby is needed. By accounting for the spin-phonon interaction, one can calculate the relaxation rates of the transitions for low spin concentrations [29]. With these relaxation rates, the inversion ratios and pump power requirements can be calculated for each operating-point of interest and for various physical temperatures.

Acknowledgments

The author acknowledges the encouragement and guidance of J. Shell, and would like to thank R. Clauss for his comments on the rough draft.

References

[1] S. Petty, "Introduction to Microwave Devices, Part VIII," in *Low Temperature Electronics*, edited by R. Kirschman, New York: IEEE Press, 1986.

[2] C. Moore and R. Clauss, "A Reflected-Wave Ruby Maser with K-Band Tuning Range and Large Instantaneous Bandwidth," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-27, pp. 249–256, March 1979.

[3] C. Moore and D. Neff, "Experimental Evaluation of Ruby at 43 GHz," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-30, pp. 2013–2015, Nov. 1982.

[4] A. Blinov and S. Peskovatskii, "Inversion Characteristics of Ruby at the Center of the Millimeter Band," *Radiofizika*, vol. 30, pp. 784–787, 1987.

[5] J. Shell and D. Neff, "A 32 GHz Reflected-Wave Maser Amplifier with Wide Instantaneous Bandwidth," in *Conference Digest, IEEE MTT-S International Microwave Symposium*, New York, May 1988.

[6] J. Orton, D. Paxman, and J. Walling, *The Solid State Maser*, Oxford: Pergamon Press Ltd., Chapter 2, Maser Materials, 1970.

[7] W. Low, "Paramagnetic Substances Suitable for Maser Operation in the Millimeter Range," in *Conference Proceedings, Symposium on Millimeter Waves*, Polytechnic Institute of Brooklyn, March 31, April 1–2, 1959.

[8] A. Abragam and B. Bleaney, *Electron Paramagnetic Resonance of Transition Ions*, New York: Dover, 1986.

[9] A. Siegman, *Microwave Solid-State Masers*, New York: McGraw-Hill, 1964.

[10] Schiff, *Quantum Mechanics*, 3rd ed., New York: McGraw-Hill, 1968.

[11] E. Schulz-Du Bois, "Paramagnetic Spectra of Substituted Sapphires–Part 1: Ruby," *Bell Syst. Tech J.*, vol. 38, pp. 271–290, 1959.

[12] J. Geusic, M. Peter, and E. Schulz-Du Bois, "Paramagnetic Resonance Spectrum of CR^{3+} in Emerald," *Bell Syst. Tech. J.*, vol. 38, pp. 291–296, 1959.

[13] R. Stahl-Brada and W. Low, "Paramagnetic Resonance Spectra of Chromium and Manganese in the Spinel Structure," *Phys. Rev.*, vol. 116, pp. 561–564, 1959.

[14] J. Carson and R. White, "Zero-Field Splitting of the Cr^{3+} Ground State of YGa and YAl Garnet," *J. Appl. Phys.*, vol. 32, p. 1787, 1961.

[15] H. Gerritsen, S. Harrison, and H. Lewis, "Chromium-Doped Titania as a Maser Material," *J. Appl. Phys.*, vol. 31, pp. 1566–1571, 1960.

[16] S. Kurtz and W. Nilsen, "Paramagnetic Resonance Spectra of Cr^{3+} in $ZnWO_4$," *Phys. Rev.*, vol. 128, pp. 1586–1588, 1962.

[17] V. Vinokurov, et al., "Paramagnetic Resonance of Trivalent Chromium in Andalusite," *Sov. Phys.–Solid State*, vol. 4, pp. 470–472, 1962.

[18] J. Carson, D. Devon, and R. Hoskins, "Paramagnetic Resonance of Cr^{3+} in Yttrium Oxide," *Phys. Rev.*, vol. 122, pp. 1141–1143, 1961.

[19] D. Carter and A. Okaya, "Electron Paramagnetic Resonance of Fe^{3+} in TiO_2 (Rutile)," *Phys. Rev.*, vol. 118, pp. 1485–1490, 1960.

[20] W. Nilsen and S. Kurtz, "Paramagnetic Resonance Spectra of Fe^{3+} in $ZnWO_4$," *Phys. Rev.*, vol. 136, pp. A262–A266, 1964.

[21] F. Holuj, J. Thyen, and N. Hedgecock, "ESR Spectra of Fe^{3+} in Single Crystals of Andalusite," *Can. J. Phys.*, vol. 44, pp. 509–523, 1966.

[22] G. Bogle and H. Symmons, "Paramagnetic Resonance of Fe^{3+} in Sapphire at Low Temperatures," *Proc. Phys. Soc.* (London), vol. 73, pp. 531–532, 1959.

[23] B. Bleaney and W. Stevens, "Paramagnetic Resonance," *Reports on Prog. Phys.*, vol. 16, pp. 108–159, 1953.

[24] R. Berwin, "Paramagnetic Energy Levels of the Ground State of Cr^{3+} in Al_2O_3 (Ruby)," *Technical Memorandum, 33-440*, Jet Propulsion Laboratory, Pasadena, California, January 1970.

[25] R. Berwin, "Energy Levels and Transition Matrix Elements of Paramagnetic Crystals for Maser Applications," *Technical Memorandum, 33-446*, Jet Propulsion Laboratory, Pasadena, California, March 1970.

[26] J. Orton, "Paramagnetic Resonance Data," *Rep. Prog. Phys.*, vol. 22, pp. 204–240, 1959.

[27] R. Berwin, R. Clauss, and E. Wiebe, "Low-Noise Receivers: Microwave Maser Development," *JPL Space Programs Summary, 37-56*, vol. II, Jet Propulsion Laboratory, Pasadena, California, March 1969.

[28] K. Standley and R. Vaughan, "Effect of Crystal-Growth Method on Electron Spin Relaxation in Ruby," *Phys. Rev.*, vol. 139, pp. A1275–A1280, 1965.

[29] V. Shakhparyan and R. Martirosyan, "Inversion Ratio Studies of Ruby in High-Intensity Magnetic Field," *Phys. Stat. Sol. (a)*, vol. 25, pp. 681–690, 1974.

[30] K. Standley and R. Vaughan, *Electron Spin Resonance Phenomena in Solids*, London: Adam Hilger LTD, 1969.

[31] C. Moore, "A K-Band Ruby Maser with 500-MHz Bandwidth," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-28, pp. 149–151, Feb. 1980.

[32] J. Pace, D. Sampson, and J. Thorp, "Spin-Lattice Relaxation Times in Ruby at 34.6 Gc/s," *Proc. Phys. Soc.*, vol. 76, p. 697, 1960.

[33] D. Mason and J. Thorp, "Influence of Crystalline Imperfections on Spin-Lattice Relaxation in Ruby," *Phys. Rev.*, vol. 157, p. 191, 1967.

[34] J. Pace, D. Sampson, and J. Thorp, "Spin-Lattice Relaxation Times in Sapphire and Chromium-doped Rutile at 34.6 Gc/s," *Proc. Phys. Soc.*, vol. 77, p. 257, 1961.

[35] P. Squire and J. Orton, "Relaxation of the Cr^{3+} Ion in Emerald," *Proc. Phys. Soc.*, vol. 88, pp. 649–657, 1966.

[36] J. Orton, A. Fruin, and J. Walling, "Spin-Lattice Relaxation of Cr^{3+} in Single Crystals of Zinc Tungstate," *Proc. Phys. Soc.*, vol. 87, pp. 703–716, 1966.

[37] M. Madan, "Spin-Lattice Relaxation Time of Fe^{3+} Ions," *Can. J. Phys.*, vol. 42, pp. 583–594, 1964.

Table 1. Promising 32-GHz operating points for ruby and other materials. The materials are arranged in order of increasing ZFS.

Operating Point	Material					
	Ruby No. 1	Ruby No. 2	Ruby No. 3	Sapphire No. 1	Sapphire No. 2	YAG No. 1
Ion	Cr	Cr	Cr	Fe	Fe	Cr
ZFS, GHz	11.4	11.4	11.4	12.1, 19.1	12.1, 19.1	15.7
No. of ionic sites	1	1	1	2	2	1
θ, ϕ, deg	54.74	90	90	90, 45	60, 30	54.74
B, kG	11.81	13.50	11.20	9.50	12.44	13.27
f_s, GHz	32.0 (32)	32.0 (21)	32.0 (32)	32.0 (32)	32.0 (54)	32.0 (32)
f_{p1}	66.2 (13)	70.3 (13)	57.6 (13)	68.6 (13)	77.3 (13)	76.2 (13)
f_{p2}	66.2 (24)	43.3 (34)	68.9 (24)	59.9 (24)	66.1 (35)	76.2 (24)
$\Delta f_s/\Delta B$, MHz/G	2.9	2.8	2.7	2.8	3.0	2.9
$\Delta f_{p1}/\Delta f_s$	1.9	2.0	2.0	2.0	1.8	1.8
$\Delta f_{p2}/\Delta f_s$	1.9	1.0	2.0	2.0	2.0	1.8
σ_s^2	1.92	1.51	1.97	6.87	5.64	1.62
σ_{p1}^2	0.05	0.02	0.03	0.08	0.09	0.13
σ_{p2}^2	0.04	1.51	0.02	0.17	0.28	0.31
I_{opt}	3.1	2.2	2.8	3.0	4.8	3.7
I_{equal}	1.1	0.7	0.9	1.0	2.0	1.4
$\Delta f_L/Q_m$, MHz	8.3	4.7	7.7	19.2	25.3	8.4

Operating Point	Material					
	YAG No. 2	YGG No. 1	YGG No. 2	Spinel No. 1	Andalusite No. 1	Andalusite No. 2
Ion	Cr	Cr	Cr	Cr	Cr	Cr
ZFS, GHz	15.7	20.9	20.9	29.7	32.0	32.0
No. of ionic sites	1	1	1	4	2	2
θ, ϕ, deg	70	54.74	70	54.74	55, 0	70, 0
B, kG	13.40	13.96	13.35	13.15	12.94	13.00
f_s, GHz	32.0 (43)	32.0 (32)	32.0 (43)	32.0 (32)	32.0 (32)	32.0 (43)
f_{p1}	49.6 (12)	81.8 (13)	53.3 (12)	75.3 (13)	74.2 (13)	47.0 (12)
f_{p2}	69.7 (24)	81.8 (24)	70.7 (24)	75.3 (24)	75.8 (24)	69.4 (24)
$\Delta f_s/\Delta B$, MHz/G	2.4	2.7	2.2	2.9	2.8	2.4
$\Delta f_{p1}/\Delta f_s$	1.2	1.9	1.3	1.8	1.9	1.1
$\Delta f_{p2}/\Delta f_s$	2.1	1.9	2.1	1.8	1.9	2.1
σ_s^2	1.25	1.50	1.01	1.66	1.63	1.27
σ_{p1}^2	1.52	0.16	1.54	0.28	0.15	1.53
σ_{p2}^2	0.31	0.47	0.55	0.13	0.31	0.31
I_{opt}	3.6	4.0	3.9	3.6	3.6	3.5
I_{equal}	1.5	1.6	1.6	1.4	1.4	1.4
$\Delta f_L/Q_m$, MHz	6.3	8.4	5.5	8.4	8.2	6.2

Table 1 (contd)

Operating Point	Material					
	Rutile No. 1	Rutile No. 2	Rutile No. 3	Rutile No. 4	$ZnWO_4$ No. 1	$ZnWO_4$ No. 2
Ion	Cr	Cr	Fe	Fe	Cr	Cr
ZFS, GHz	43.3	43.3	43.3, 81.3	43.3, 81.3	51.6	51.6
No. of ionic sites	2	2	2	2	1	1
θ, ϕ, deg	45, 0	54.74, 45	52.55, 40	71.12, 70	40, 90	50, 90
B, kG	12.78	14.06	9.35	11.71	9.78	13.10
f_s, GHz	32.0 (43)	32.0 (32)	32.0 (32)	32.0 (43)	32.0 (21)	32.0 (32)
f_{p1}	56.0 (12)	82.6 (13)	78.9 (13)	71.8 (12)	54.1 (13)	75.1 (13)
f_{p2}	65.4 (24)	82.6 (24)	73.3 (24)	81.4 (24)	54.9 (34)	88.4 (24)
$\Delta f_s/\Delta B$, MHz/G	2.0	2.6	2.8	3.7	2.3	1.9
$\Delta f_{p1}/\Delta f_s$	1.5	1.9	2.1	0.9	1.1	2.2
$\Delta f_{p2}/\Delta f_s$	2.2	1.9	2.1	2.1	1.8	2.6
σ_s^2	1.57	1.37	3.80	3.57	1.68	1.37
σ_{p1}^2	1.18	0.48	0.67	3.31	0.53	0.31
σ_{p2}^2	0.37	0.30	1.55	1.01	0.57	0.61
I_{opt}	3.7	4.1	3.7	5.7	2.1	3.8
I_{equal}	1.4	1.6	1.4	2.6	0.5	1.5
$\Delta f_L/Q_m$, MHz	8.1	7.9	13.1	19.0	4.9	7.3

Operating Point	Material					
	Emerald No. 1	Emerald No. 2	$ZnWO_4$ No. 3	$ZnWO_4$ No. 4	Y_2O_3 No. 1	Andalusite No. 3
Ion	Cr	Cr	Fe	Fe	Cr	Fe
ZFS, GHz	53.5	53.5	61.0, 76.9	61.0, 76.9	72.7	112.6, 225.2
No. of ionic sites	1	1	1	1	4	2
θ, ϕ, deg	40, 0	54.74	90, 45	90, 45	40	30, 0
B, kG	7.95	14.04	8.59	5.28	9.14	6.63
f_s, GHz	32.0 (43)	32.0 (32)	32.0 (54)	32.0 (54)	32.0 (21)	32.0 (21)
f_{p1}	48.4 (12)	86.1 (13)	66.0 (13)	67.9 (13)	69.1 (13)	107.3 (13)
f_{p2}	50.8 (24)	86.1 (24)	67.4 (35)	57.6 (35)	88.3 (24)	–
$\Delta f_s/\Delta B$, MHz/G	3.5	2.2	2.1	−2.7	−3.32	4.78
$\Delta f_{p1}/\Delta f_s$	1.5	2.2	−0.1	0.4	−0.16	−0.07
$\Delta f_{p2}/\Delta f_s$	0.3	2.2	2.1	−0.5	−0.72	–
σ_s^2	1.54	1.38	5.52	3.24	0.76	3.55
σ_{p1}^2	0.13	0.66	3.01	3.17	0.73	2.32
σ_{p2}^2	0.59	0.19	2.91	3.51	1.31	–
I_{opt}	2.6	4.3	4.3	3.8	3.6	2.0
I_{equal}	0.8	1.7	1.8	1.5	1.4	0.5
$\Delta f_L/Q_m$, MHz	5.6	8.3	22.2	11.5	3.8	6.6

Appendix
Table of Relaxation Times

This Appendix contains a table of relaxation times (T_1) for various materials at 4.2 K. The paramagnetic ion concentration, transition frequency, transition, and orientation are also shown. In addition, the literature references are included. Note that the table does not mention either the crystal growth process or the relaxation time measurement technique, both of which may significantly alter the reported relaxations times. Hence, these relaxation times could easily be an order of magnitude or more different from what one might measure, and should not be used in calculations unless verified.

Table A-1. Measured relaxation times, T_1, for various materials at physical temperature T = 4.2° K

Material	Ionic concentration, atomic percent	Frequency, GHz	Orientation, deg (θ, ϕ)	Transition	T_1, msec	Reference
Ruby	0.03	34.6	90	3–4	21	[32]
				2–3	16	
				1–2	22	
				2–4	54	
				1–3	56	
	0.013	35	90	2–3	15.5	[33]
	0.052			2–3	17.5	
Sapphire: Fe	0.03	34.6	90, 0	4–5	1.8	[34]
				3–4	1.6	
				2–3	1.5	
				1–2	2.0	
Emerald	4.9×10^{19} ions/cm^3	9.3	0	3–4	9	[35]
			90	1–2	8	
				3–4	11	
$ZnWO_4$: Cr	0.005–0.3	9.2	90, 90	1–2	≈1.5	[36]
	0.018–0.72	33	90, 90	1–2	≈0.5	
	0.08	X-band	?	5–6	~0.3	[a]
				1–2	~1	
Rutile: Cr	0.07	34.6	90, 0[b]	1–2 (Site A)	4.5	[34]
				3–4 (Site A)	2.5	
				2–3 (Site B)	2.5	
				2–4 (Site B)	2.1	
Rutile: Fe	0.01–0.02	9.4	0, 0	1–2	≈2	[37]
				3–4	≈2	

[a] J. Orton, private communication with K. Standley and R. Vaughan.
[b] For this orientation the ionic sites are inequivalent.

32-GHz Cryogenically Cooled HEMT Low-Noise Amplifiers

J. J. Bautista and G. G. Ortiz
Radio Frequency and Microwave Subsystems Section

K. H. G. Duh, W. F. Kopp, P. Ho, P. C. Chao, M. Y. Kao,
P. M. Smith, and J. M. Ballingall
GE Electronics Laboratory

The cryogenic noise temperature performance of a two-stage and a three-stage 32-GHz HEMT amplifier has been evaluated. The amplifiers employ 0.25-μm conventional AlGaAs/GaAs HEMT devices, hybrid matching input and output microstrip circuits, and a cryogenically stable dc biasing network. The noise temperature measurements were performed in the frequency range of 31 to 33 GHz over a physical temperature range of 300 K down to 12 K. Across the measurement band, the amplifiers displayed a broadband response, and the noise temperature was observed to decrease by a factor of 10 in cooling from 300 K to 15 K. The lowest noise temperature measured for the two-stage amplifier at 32 GHz was 35 K with an associated gain of 16.5 dB, while the three-stage amplifier measured 39 K with an associated gain of 26 dB. It was further observed that both amplifiers were insensitive to light.

I. Introduction

Traditionally, the extraordinarily sensitive receiver systems operated by the Jet Propulsion Laboratory's Deep Space Network (DSN) have employed ruby masers as the low-noise front-end amplifiers. The rapid advances recently achieved by cryogenically cooled high-electron-mobility transistor (HEMT) low-noise amplifiers (LNAs) in the 1- to 10-GHz range are approaching maser amplifier performance [1], [2]. In order to address its future spacecraft navigation, telemetry, radar, and radio science needs, the DSN is investigating both maser [3] and HEMT amplifiers for its 32-GHz downlink capability. This report describes the noise temperature performance of the 32-GHz HEMT LNAs.

II. HEMT Device

Since one of the primary functions of the LNA is to minimize the receiver system noise temperature, the characterization and selection of HEMT devices is critical to the LNA's performance. The selection of the 0.25-μm gate length conventional AlGaAs/GaAs HEMTs was based on their previously demonstrated reliability and exceptionally high gain and low noise characteristics [4], [5].

The devices were fabricated on selectively doped AlGaAs/GaAs heterostructures grown by molecular beam epitaxy (MBE) with a Varian GEN II system on a 3-inch-diameter GaAs substrate. The details of the material growth conditions

are discussed elsewhere [6]. Figure 1 schematically illustrates the cross section of the HEMT device. The HEMT wafer exhibited a sheet carrier density of $8.1 \times 10^{12}/cm^2$ with a mobility of more than 75,000 $cm^2/V \cdot sec$ at 77 K. All levels were defined by electron beam lithography, and the T-shape gates were fabricated using the PMMA/P(MMA-MMA)/PMMA tri-layer resist technique [7] to achieve a low series gate resistance.

For low-noise performance at cryogenic temperatures, the HEMT device must exhibit good pinch-off characteristics and high transconductance, g_m. Good pinch-off characteristics are achieved by strong confinement of the charge carriers to the channel region with a sharp interface of high quality and a large conduction band discontinuity. An enhanced g_m at the operating bias is obtained by a judicious choice of doping concentration and space layer thickness [8]. An Al mole fraction of approximately 30% is required for a large conduction band discontinuity, while the high g_m is achieved with a 4-nm spacer layer and a doping concentration of approximately 2×10^{18} dopant atoms/cm^3. Although these values will result in a high-performance room-temperature device, at physical temperatures below 150 K the device will suffer from I-V collapse [9] and exhibit the persistent photoconductivity effect associated with the presence of deep donor traps (called DX centers). In order to obtain excellent device performance at cryogenic temperatures and to eliminate light sensitivity, previous work [1], [8] has demonstrated that the Al composition must not exceed 23% and the doping concentration must be approximately $10^{18}/cm^3$.

The data shown in Table 1, comparing two HEMTs with the same Al mole fraction (23%) but different doping concentrations in the n-AlGaAs layer, serves to illustrate the difference between low-temperature and room-temperature device optimization. Device A has an n-AlGaAs doping concentration of $10^{18}/cm^3$, while that of B is twice as high. As expected, device B exhibited a higher g_m and associated gain than device A, with approximately the same noise figure for both devices at 300 K. However, at 13 K and 8.5 GHz, device B exhibited a minimum noise temperature of 13.1 K, while device A yielded a value of 5.3 K.

III. Amplifier Design and Circuit

Both LNAs were designed to achieve the best room-temperature low-noise performance based on the measured room-temperature device parameters. Following construction and room-temperature optimization, the LNAs are then biased for lowest noise performance at cryogenic temperatures.

The device gate width of 75 μm selected for this work was determined by the tradeoffs associated with optimum impedance matching, circuit bandwidth, intermodulation distortion, power handling capability, and power dissipation. Figure 2 shows a photograph of the 32-GHz hybrid two-stage HEMT LNA package. The input and output ports utilize a broadband WR28-to-stepped ridge waveguide-to-microstrip transition. Figure 3 shows the insertion loss and return loss of a stepped ridge fixture that consists of two stepped-ridge transitions connected back-to-back with a microstrip 50-ohm line 0.5 in. long. The input and output matching networks were designed based on the device equivalent circuit values obtained from fitting measured S-parameters at the low-noise bias condition to the model from 2 to 20 GHz. Figure 4 shows the topology used for the 0.25-μm HEMT equivalent circuit model. Input, output, and interstage matching circuits were designed on 10-mil quartz substrate with TaN thin-film resistors and TiWAu metallization. A schematic diagram of the two-stage hybrid HEMT LNA is shown in Fig. 5. The edge-coupled symmetric microstrip dc blocking transmission line also served as a bandpass filter, improving the out-of-band stability. As shown in Fig. 6, the three-stage LNA is constructed from the two-stage LNA by the insertion of another interstage matching circuit.

The LNA fixture (OFHC copper) and dc bias circuits [10] are designed for operation at cryogenic temperatures. Diode protection was included in both the gate and drain bias circuits. LEDs were mounted on the cover of the fixture above each of the HEMTs for the purpose of examining their light sensitivity at cryogenic temperatures. All of the stages use devices from the same wafer.

IV. Measurement Results

The LNAs were first measured at room temperature with the devices biased for lowest noise at room temperature and then biased for lowest noise performance at cryogenic temperatures. The LNA room-temperature broadband noise figure and gain for the two- and three-stage LNAs are shown in Figs. 7 and 8, respectively. Both LNAs exhibited an average noise figure of approximately 2 dB from 28 to 36 GHz. From 29 to 34 GHz, the gain measured approximately 17 dB and 24 dB for the two-stage and three-stage LNA, respectively. The addition of an external isolator only slightly degraded the gain and noise figure by 0.3 dB.

With the devices biased for lowest noise at cryogenic temperature (12 K), the noise temperature (referenced at the room-temperature input waveguide flange) of both LNAs was observed to decrease nearly quadratically as a function of physical temperature as they cooled from 300 K to 12 K. (See Fig. 9 for a diagram of the closed-cycle refrigerator and measurement system.) The noise temperature of the two-stage LNA decreased from 350 K at ambient to 35 K at

14.5 K, while the three-stage LNA decreased from 400 K to 41 K at 12.5 K (Figs. 10 and 11). Figures 12 and 13 show the cryogenic noise temperature and gain response from 31 to 33 GHz, along with bias settings for the two-stage and three-stage LNA, respectively. At 32 GHz, the two-stage LNA noise temperature measured 35 K, with an associated gain of 16.5 dB, at a physical temperature of 14 K, while the three-stage LNA yielded a value of 41 K with a 26.0-dB associated gain. It is also noted that the three-stage LNA displayed an almost flat noise temperature response across the measurement band, with a minimum noise temperature of 39 K at 32 GHz, while the two-stage LNA displayed a noise temperature response decreasing monotonically from 31 to 33 GHz, with a minimum noise temperature of 31 K at 33 GHz.

It was further observed that both amplifiers did not show a persistent photoconductivity effect. That is, it was found that these devices can be cooled with or without illumination and/or dc bias, without any observable effect on the cryogenic low-noise performance.

V. Conclusion

Cryogenic coolable state-of-the-art 32-GHz HEMT LNAs have been demonstrated using 0.25-μm AlGaAs/GaAs HEMTs. The results clearly demonstrate their potential to meet the future need for extremely low-noise receivers for applications such as the DSN. Further advances in HEMT technology [12] promise to lead to improved performance at all frequencies and make possible the development of amplifiers operating at frequencies up to 94 GHz.

Currently, the DSN relies on maser amplifiers in order to provide the best possible telemetry support for deep space missions. These systems require a complex and expensive cryogenic system operating at 4.5 K. Since HEMT LNAs require less cooling power and operate at a higher physical temperature (12 K), they can be operated at less cost with a more reliable refrigeration system. The lower cost of HEMT LNAs will lead to greater frequency coverage and the economic realization of multiple-element cryogenic array feed systems.

Acknowledgments

The authors wish to thank M. Pospieszalski for providing X-band cryogenic HEMT data, and J. Merrill for the waveguide transition design. The authors would also like to acknowledge the support of A. A. Jabra, D. Neff, J. Bowen, and D. Norris. GE Electronics Laboratory's 32-GHz cryogenic low-noise HEMT development was supported by JPL under Contract No. 957352.

References

[1] M. W. Pospieszalski, S. Weinreb, P. C. Chao, U. K. Mishra, S. C. Palmateer, P. M. Smith, and J. C. M. Hwang, "Noise Parameters and Light Sensitivity of Low-Noise High-Electron-Mobility Transistors," *IEEE Trans. Electron Devices*, vol. ED-33, pp. 218–223, 1986.

[2] S. Weinreb, M. W. Pospieszalski, and R. Norrod, "Cryogenic, HEMT, Low-Noise Receivers for 1.3 to 43 GHz Range," *IEEE MTT-S Digest*, pp. 945–948, 1988.

[3] J. Shell and D. Neff, "A 32-GHz Reflected Wave Maser Amplifier with Wide Instantaneous Bandwidth," *IEEE MTT-S Digest*, pp. 789–792, 1988.

[4] K. H. G. Duh, P. C. Chao, P. M. Smith, L. F. Lester, B. R. Lee, J. M. Ballingall, and M. Y Kao, "Millimeter-Wave Low-Noise HEMT Amplifiers," *IEEE MTT-S Digest*, pp. 923–926, 1988.

[5] P. M. Smith, P. C. Chao, K. H. G. Duh, L. F. Lester, B. R. Lee, J. M. Ballingall, and M. Y. Kao, "Advances in HEMT Technology and Applications," *IEEE MTT-S Digest*, pp. 749–752, 1987.

[6] S. C. Palmateer, P. A. Maki, W. Katz, A. R. Calawa, J. C. M. Hwang, and L. F. Eastman, "The influence of V:III flux ratio on unintentional impurity incorporation during molecular beam epitaxial growth," in *Proc. Gallium Arsenide and Related Compounds 1984* (Inst. Phys. Conf. Series 74), pp. 217–222, 1985.

[7] P. C. Chao, P. M. Smith, S. C. Palmateer, and J. C. M. Hwang, "Electron-Beam Fabrication of GaAs Low-Noise MESFETs Using a New Tri-Layer Resist Technique," *IEEE Trans. Electron Devices*, vol. ED-22, pp. 1042–1046, 1985.

[8] K. H. G. Duh, M. W. Pospieszalski, W. F. Kopp, P. Ho, A. A. Jabra, P. C. Chao, P. M. Smith, L. F. Lester, J. M. Ballingall, and S. Weinreb, "Ultra-Low-Noise Cryogenic High-Electron Mobility Transistors," *IEEE Trans. Electron Devices*, vol. ED-35, pp. 249–256, 1988.

[9] A. Kastarsky and R. A. Klein, "On the low temperature degradation of AlGaAs/GaAs modulation doped field-effect transistors," *IEEE Trans. Electron Devices*, vol. ED-33, pp. 414–423, 1986.

[10] S. Weinreb and R. Harris, "A 23-GHz Coolable FET Amplifier," NRAO Internal Report.

[11] P. C. Chao, P. M. Smith, K. H. G. Duh, J. M. Ballingall, L. F. Lester, B. R. Lee, A. A. Jabra, and R. C. Tiberio, "High Performance 0.1 μm Gate-Length Planar-Doped HEMTs," 1987 International Electron Devices Meeting, Washington, D. C., Dec. 1987, Paper 17.1, pp. 410–413.

[12] P. Ho, P. C. Chao, K. H. G. Duh, A. A. Jabra, J. M. Ballingall, and P. M. Smith, "Extremely High Gain, Low Noise InAlAs/InGaAs HEMTs Grown by Molecular Beam Epitaxy," to be published in *1988 IEDM Technical Digest*.

Table 1. Performance comparison of two types of conventional AlGaAs/GaAs HEMTs

Ambient Temp., K	Freq., GHz	Performance parameter	Type A (10^{18}/cm^3)	Type B (2×10^{18}/cm^3)
300		g_m (mS/mm)	380	450
300	8	Noise figure (dB)	0.4	–
300	8	Associated gain (dB)	15.2	–
300	18	Noise figure (dB)	0.7	0.7
300	18	Assoc. gain (dB)	11.5	15.0
300	32	Noise figure (dB)	1.3	1.2
300	32	Assoc. gain (dB)	7.5	10.0
13	8.5	Noise temp. (K)	5.3	13.1
13	8.5	Assoc. gain (dB)	13.9	14.5

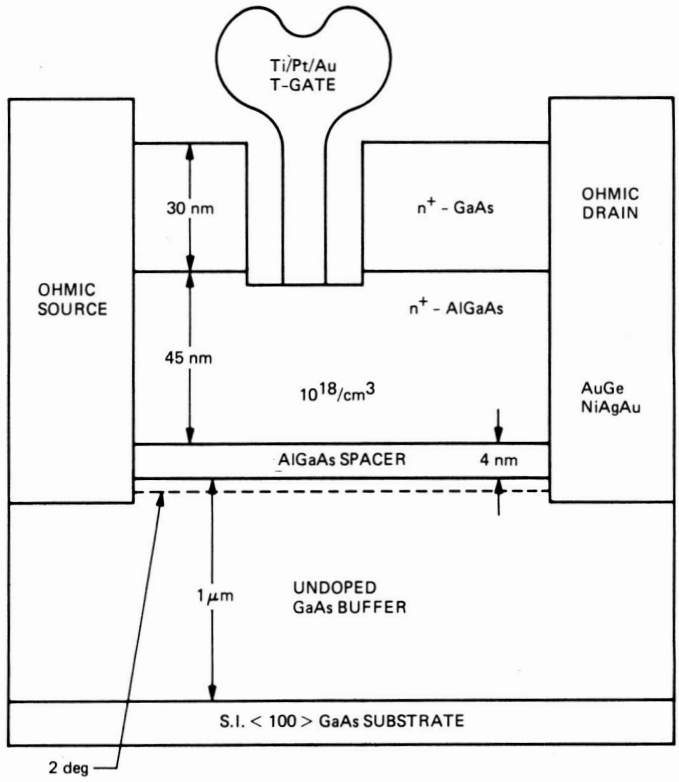


Fig. 1. Cross section of the 0.25-μm T-gate HEMT.

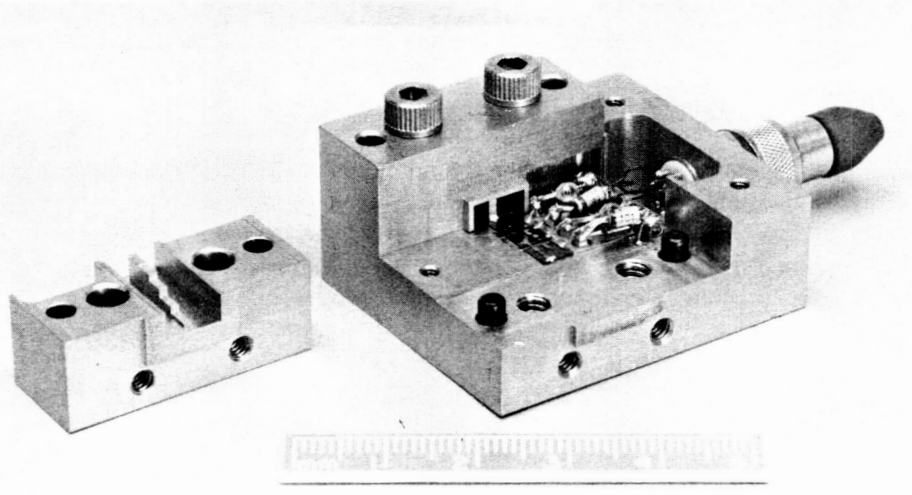


Fig. 2. Ka-band two-stage HEMT LNA.

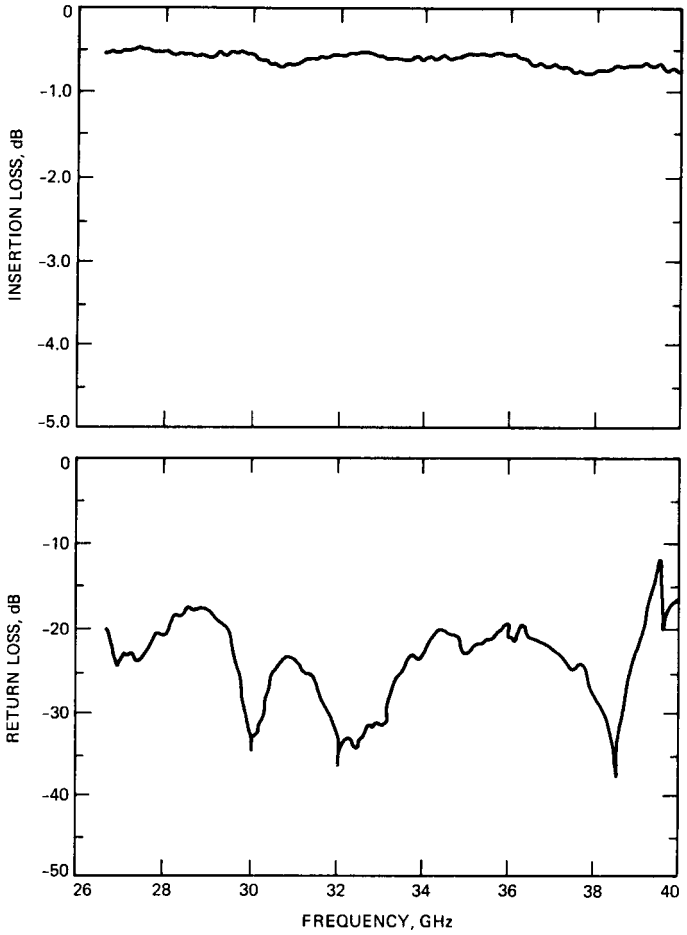


Fig. 3. Measured performance of Ka-band stepped ridge fixture.

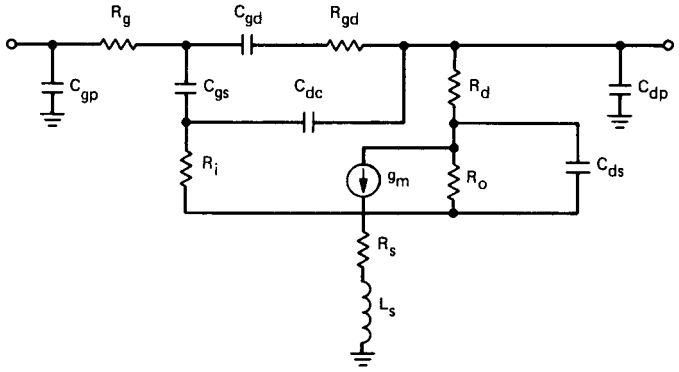


Fig. 4. Equivalent circuit model of 0.25-μm AlGaAs/GaAs HEMT.

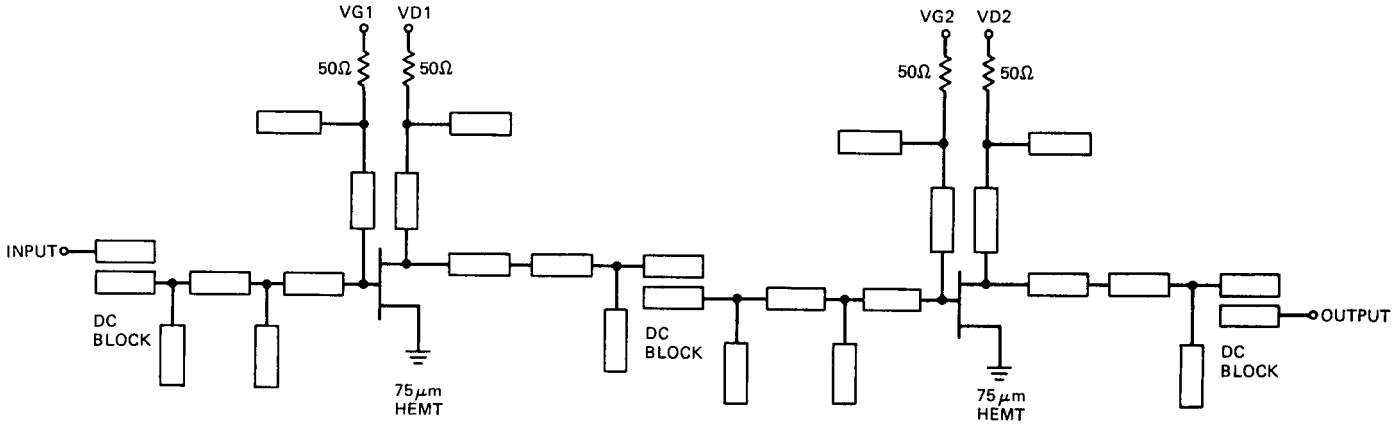


Fig. 5. Schematic diagram of two-stage hybrid LNA using 0.25x75-μm HEMTs.

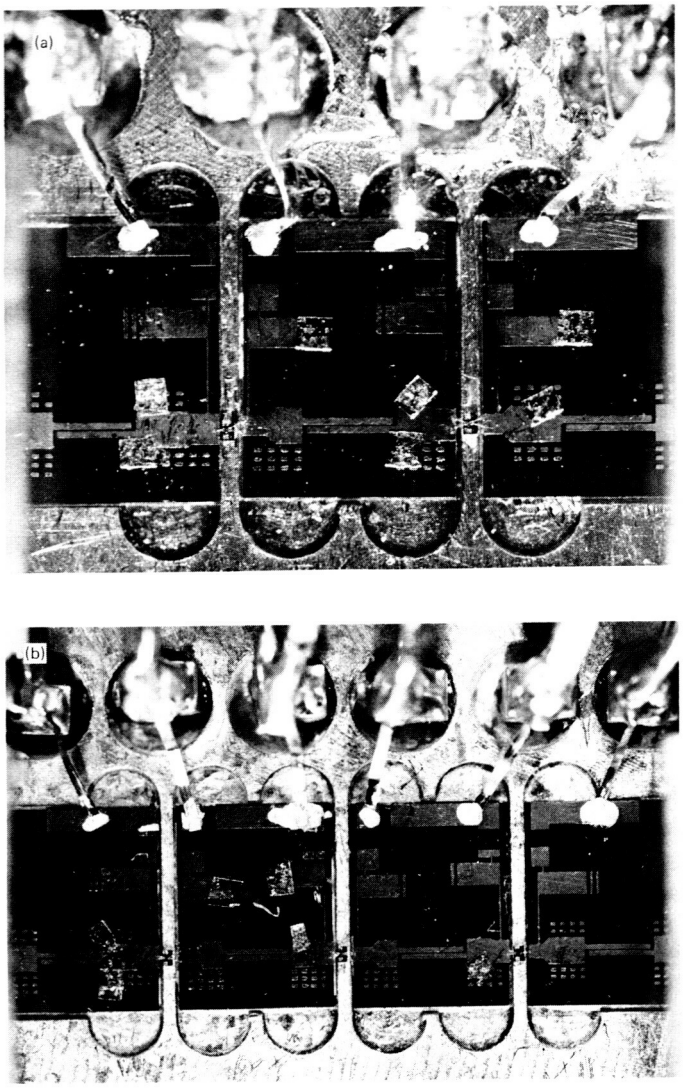


Fig. 6. Two-stage (a) and three-stage (b) HEMT LNA microstrip matching circuits.

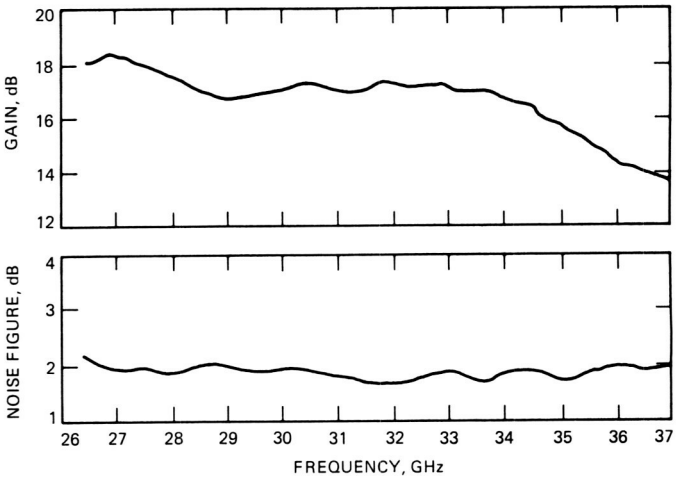


Fig. 7. Two-stage HEMT LNA room-temperature broadband response.

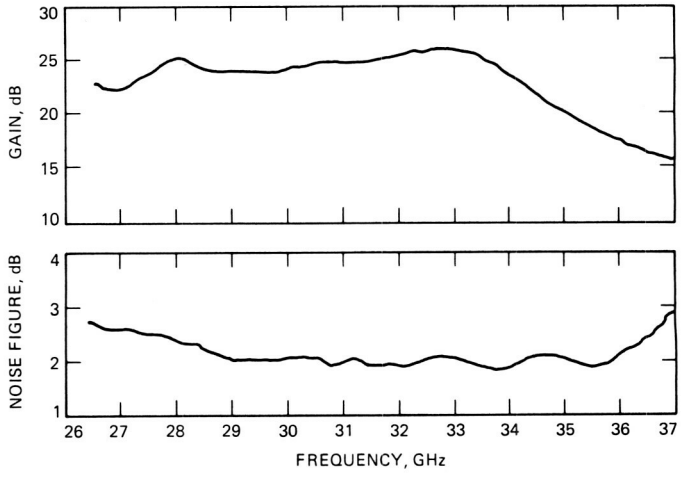


Fig. 8. Three-stage HEMT LNA room-temperature broadband response.

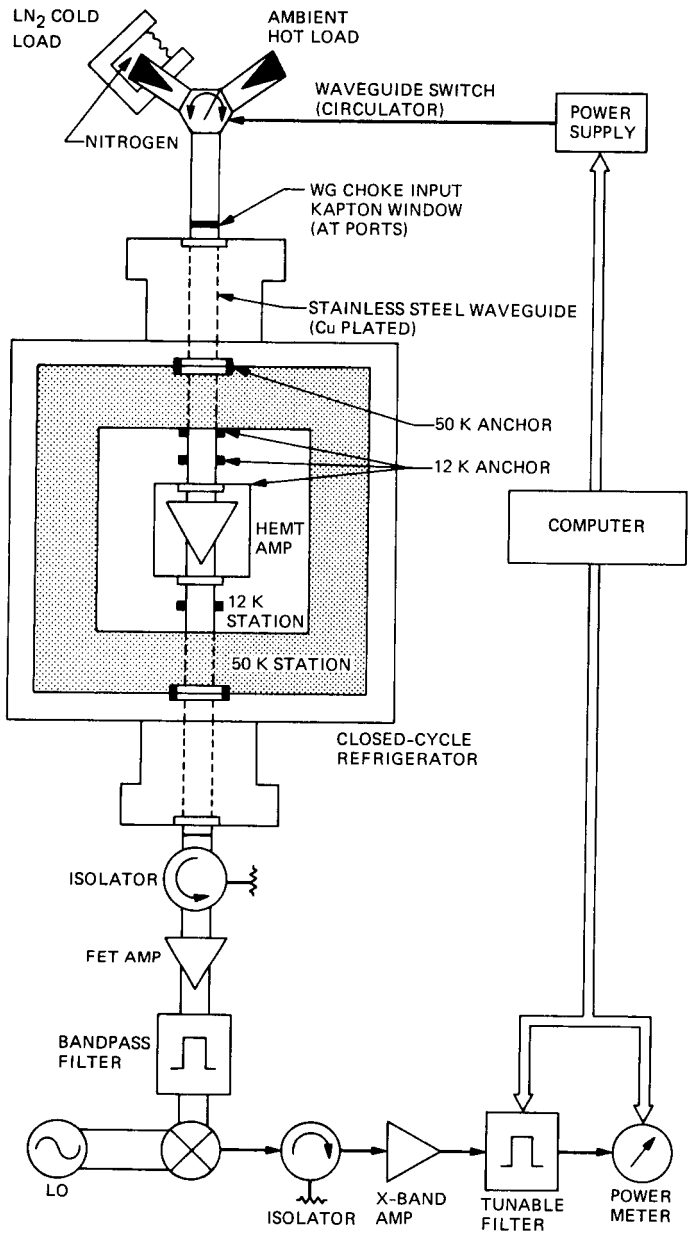


Fig. 9. 32-GHz cryogenic noise temperature measurement system.

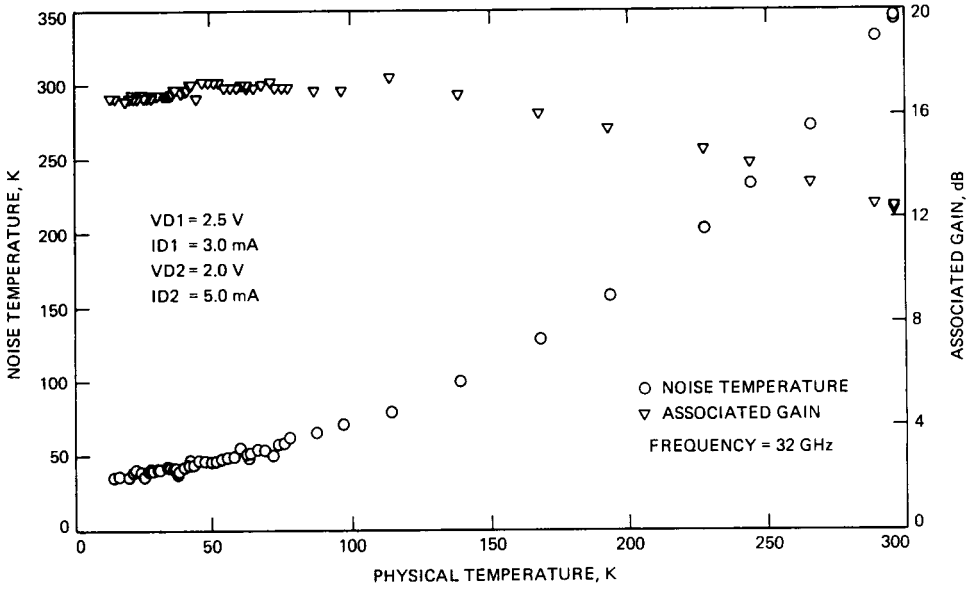


Fig. 10. Two-stage HEMT LNA noise temperature and gain cooling curve.

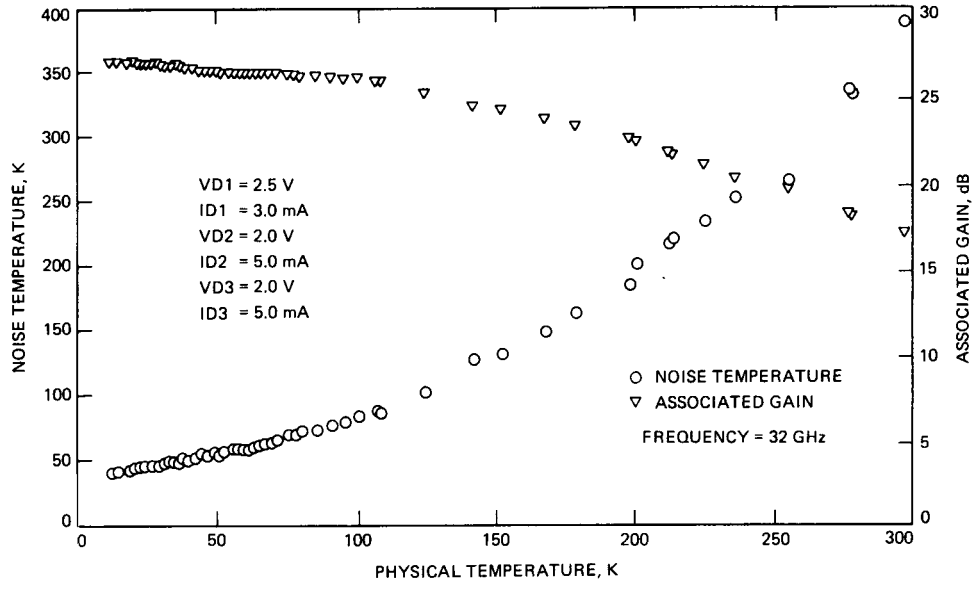


Fig. 11. Three-stage HEMT LNA noise temperature and gain cooling curve.

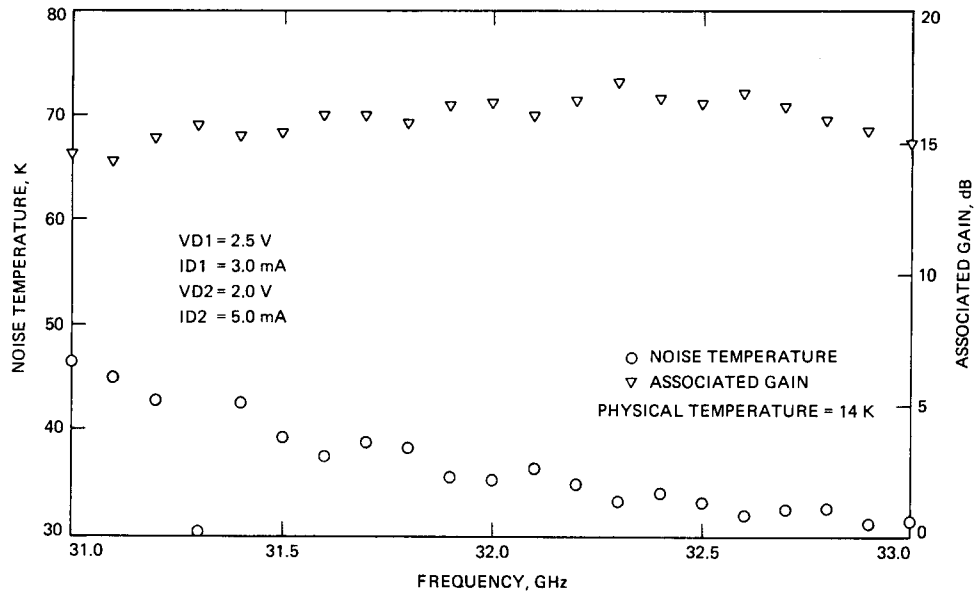


Fig. 12. Two-stage HEMT LNA at 14-K noise temperature and gain.

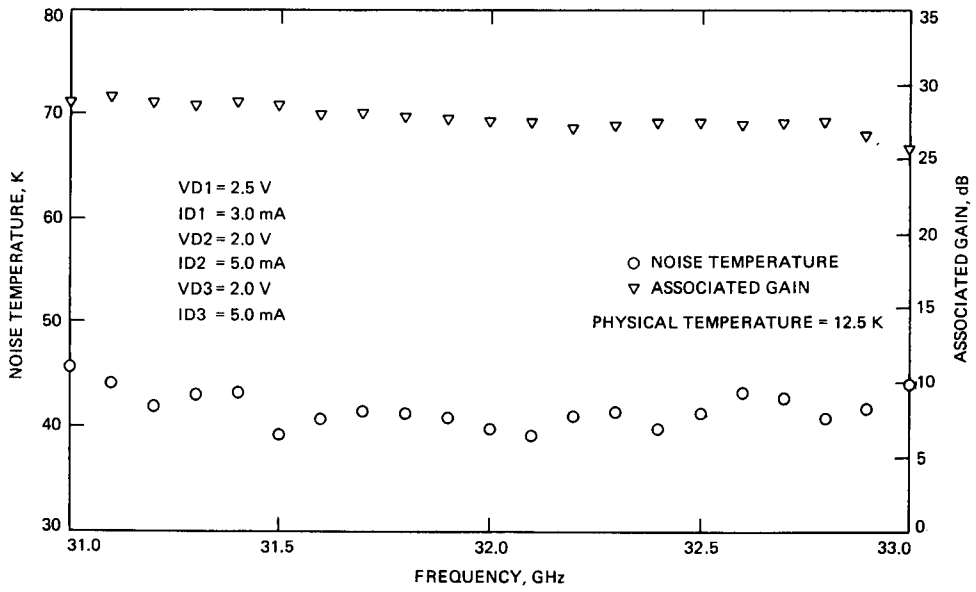


Fig. 13. Three-stage HEMT LNA at 12.5-K noise temperature and gain.

Cross-Guide Coupler Modeling and Design

J. Chen
Radio Frequency and Microwave Subsystems Section

This report describes modeling of cross-guide couplers based on the theory of equivalent electric and magnetic dipoles of an aperture. Additional correction factors due to nonzero wall thickness and large aperture are also included in this analysis. Comparisons of the measured and calculated results are presented for cross-guide couplers with circular or cross-shaped coupling apertures. A cross-guide coupler was designed as a component of the C-band feed to support the Phobos mission.

I. Introduction

In this article, the performance of a directional coupler is specified by two parameters: coupling and directivity [1]. The circular aperture and cross-shaped aperture are chosen because the circular aperture is the most well-studied aperture shape and the cross-shaped aperture is frequently used for cross-guide couplers.

A program was developed to calculate the coupling and the directivity of circular or cross-shaped aperture couplers. Either one aperture (Fig. 1) or two symmetrically spaced apertures (Fig. 2) may be used for directional coupling. The angle between two waveguides is arbitrary for one-aperture couplers, but it is 90 degrees for two-aperture couplers. The apertures may be moved along the diagonal line in the common broad wall of the waveguides. Also, the distance between two apertures located on the diagonal line varies. The coupling and the directivity for directional couplers may be calculated over a band of frequencies.

II. Theory

A. Basic Formulas

The formulas for cross-guide couplers with off-centered apertures are derived in the Appendix. The radiation amplitude in the secondary waveguide for the coupled port (B^+) and the isolated port (B^-) for the coupler of Fig. 1 is given by

$$B^+(p,m,d,\theta) = B_1(p,d) + B_3(m,d)\cos\theta + G(m,d)\sin\theta$$

$$B^-(p,m,d,\theta) = B_1(p,d) + B_4(m,d)\cos\theta$$

where θ is the angle between two waveguides, d is the distance from the center of the aperture to the waveguide wall, and p and m are the electric polarizability and magnetic polarizability of the aperture.

$$p = p_0 \cdot FE \cdot TANE$$

$$m = m_0 \cdot FM \cdot TANM$$

Here, p_0 and m_0 are the electric polarizability and magnetic polarizability of a small aperture with zero wall thickness. They are constants that depend on the shape and the size of the aperture. For a circular aperture, $p_0 = \frac{2}{3}r^3$, $m_0 = \frac{4}{3}r^3$, where r is the radius of the aperture. Values for p_0 and m_0 for a cross-shaped aperture are given in [2]. *FE*, *FM*, *TANE*, and *TANM* are defined below.

B. Wall Thickness Factor

The finite wall thickness of an aperture has the effect of reducing the coupling. The effect of finite thickness is taken into account by treating the aperture as a finite length of waveguide beyond cutoff. $FE(t)$ and $FM(t)$ represent the electric attenuation and magnetic attenuation in the aperture of thickness t and are given by [3]

$$FE(t) = e^{-2\pi\left[\left(\frac{1}{\lambda_{c1}}\right)^2 - \frac{1}{\lambda^2}\right]^{1/2} t \cdot AE}$$

$$FM(t) = e^{-2\pi\left[\left(\frac{1}{\lambda_{c2}}\right)^2 - \frac{1}{\lambda^2}\right]^{1/2} t \cdot AM}$$

where λ_{c1} = the cutoff wavelength of TM_{01} mode of aperture waveguide, λ_{c2} = the cutoff wavelength of TE_{11} mode of aperture waveguide, $\lambda_{c1} = 2.6127r$, $\lambda_{c2} = 3.4126r$ for a circular waveguide, while λ_{c1} and λ_{c2} of a cross-shaped waveguide are given in [4] and [5].

The additional factors AE and AM are effective wall thickness coefficients. For a circular aperture with radius r and thickness t [6], [7]

$$AE = 1.0103 + 0.0579\frac{r}{t} \qquad t/r > 0.2$$

$$AM = 1.0064 + 0.0819\frac{r}{t}$$

$$AE = 1.1091 - 0.0082268\frac{r}{t} \qquad t/r \leqslant 0.2$$

$$AM = 1.4273 - 0.0023284\frac{r}{t}$$

The AE and AM of a cross-shaped aperture are determined experimentally.

C. Large Aperture Factor

An infinitely thin aperture actually has an unlimited number of resonances. For a large aperture, the following frequency correction factors are needed [8]:

$$TANE = \frac{2f_{01}}{\pi f} \tan\frac{\pi f}{2f_{01}}$$

$$TANM = \frac{2f_{02}}{\pi f} \tan\frac{\pi f}{2f_{02}}$$

The resonant frequency is approximately equal to the cutoff frequency of a waveguide having the same cross-sectional shape and size as the aperture. Therefore, f_{01} may be replaced by f_{c1} and f_{02} may be replaced by f_{c2}.

$$f_{c1} = \frac{c}{\lambda_{c1}}$$

$$f_{c2} = \frac{c}{\lambda_{c2}}$$

III. Results

A. One-Circular-Aperture Coupler with θ = 45 Degrees

An off-centered circular-aperture coupler with adjustable θ has been fabricated. The dimensions of the WR112 coupler are 0.17-inch aperture radius, 0.128-inch thickness, and 0.283-inch distance from center of aperture to waveguide wall. For θ = 45 degrees, the calculated and the measured coupling and directivity at frequencies from 7 to 9 GHz are shown in Figs. 3 and 4, respectively. The calculated coupling is 0.5 to 0.7 dB higher and the directivity is 0.9 to 1.1 dB lower than measured. The results show good agreement between the coupler model and the experiment.

B. Two-Circular-Aperture Cross-Guide Coupler

A two-aperture cross-guide coupler using circular apertures was designed based on the preceding theory. The final coupler design had the following dimensions: WR125, 0.13-inch aperture radius, 0.05-inch thickness, and 0.3125-inch distance from center of aperture to waveguide wall. The measured coupling is 0.2 to 0.6 dB lower than calculated, while the directivity is 1.0 to 1.4 dB higher than expected at frequencies from 7 to 9 GHz (Figs. 5 and 6). In this case, and in general, the coupling is predicted more accurately than the directivity. The coupler computer program provides a pessimistic value of directivity.

C. Two-Cross-Shaped-Aperture Cross-Guide Coupler

As an example of a design using cross-shaped apertures, a C-band coupler meeting the following requirements was designed.

Coupling: −30 ± 1 dB

Directivity: 20 dB minimum

Waveguide: WR187

Frequency: 4.96–5.06 GHz

The final design uses two symmetric cross-shaped apertures of 0.662-inch length, 0.115-inch width, 0.05-inch thickness, and located at a 0.468-inch distance from center of aperture to waveguide wall (Fig. 7). $AE = 1.36$ and $AM = 1.53$ were determined from a previous WR125 30-dB coupler experiment. The coupler is expected to have approximately -30-dB coupling and 25.6-dB directivity according to the coupler computer program.

The measured and computed results are shown in Figs. 8 and 9. In this case, the experimental coupling is approximately 0.4 dB lower and the directivity is 0.8 to 1.5 dB higher than calculated. Figures 8 and 9 show that the directional coupler meets the design requirements.

IV. Conclusion

A brief description of a cross-guide coupler model was presented. Three specific examples have demonstrated good agreement between experiment and theory. In most cases, a suitable coupler can be designed using the simple theory presented in this report. Further directional coupler study is required to improve the prediction of directivity. Coupling between apertures and the nonuniform field over the aperture could be included in the model to obtain higher accuracy.

Acknowledgment

The author would like to acknowledge P. Stanton for the stimulating discussions and for his help in experiment design and testing.

References

[1] R. E. Collin, *Foundations for Microwave Engineering*, New York: McGraw-Hill Physical and Quantum Electronics Series, sec. 4.11 and sec. 6.4, 1966.

[2] G. Matthaei, L. Young, and E. M. T. Jones, *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, New York: McGraw-Hill Book Company, p. 233, p. 235, 1946.

[3] C. G. Montgomery, *Technique of Microwave Measurements*, New York: McGraw-Hill Book Company, sec. 14.3, 1947.

[4] F.-L. C. Lin, "Modal Characteristics of Crossed Rectangular Waveguides," *IEEE Trans. Microwave Theory and Technique*, vol. MTT-25, no. 9, pp. 756–763, September 1977.

[5] C. C. Tham, "Modes and Cutoff Frequencies of Crossed Rectangular Waveguides," *IEEE Trans. Microwave Theory and Technique*, vol. MTT-25, no. 7, pp. 585–588, July 1977.

[6] N. A. McDonald, "Electric and Magnetic Coupling through Small Aperture in Shield Wall of Any Thickness," *IEEE Trans. Microwave Theory and Technique*, vol. MTT-20, no. 10, pp. 689–695, October 1972.

[7] R. Levy, "Improved Single and Multiaperture Waveguide Coupling Theory, Including Explanation of Mutual Interactions," *IEEE Trans. Microwave Theory and Technique*, vol. MTT-28, no. 4, pp. 331–338, April 1980.

[8] S. B. Cohn, "Microwave Coupling by Large Aperture," *Proceedings of the I.R.E.*, vol. 66, pp. 696–699, June 1952.

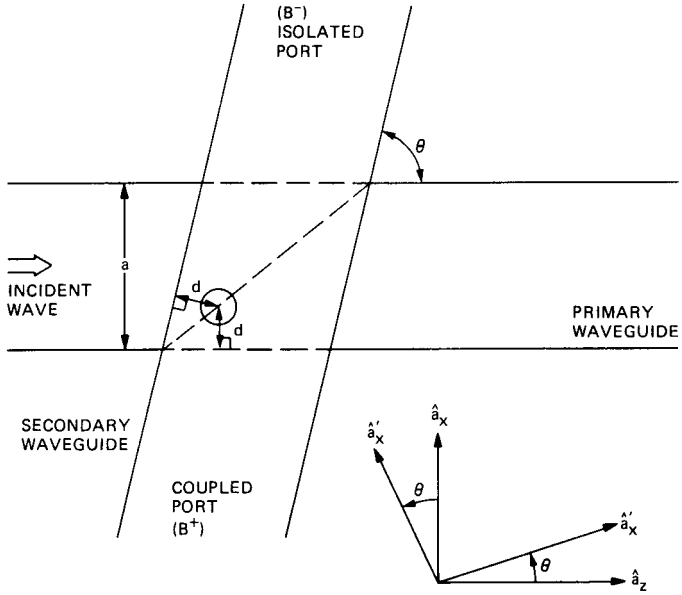


Fig. 1. Configuration of single-aperture coupler.

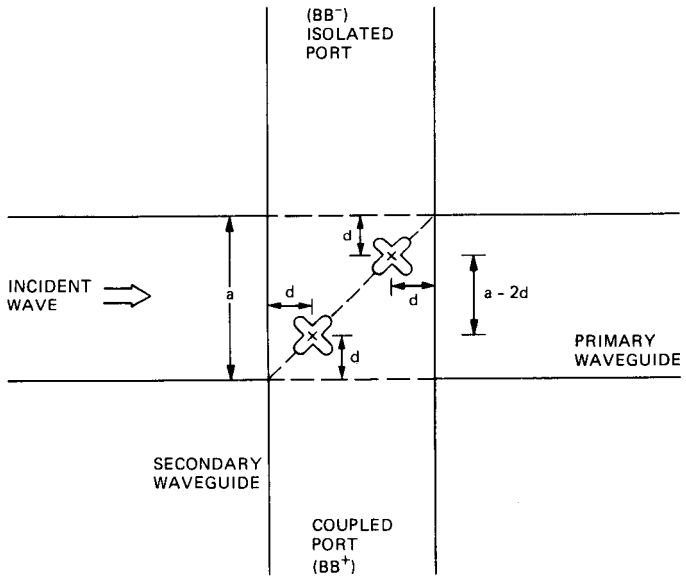


Fig. 2. Configuration of two-aperture cross-guide coupler.

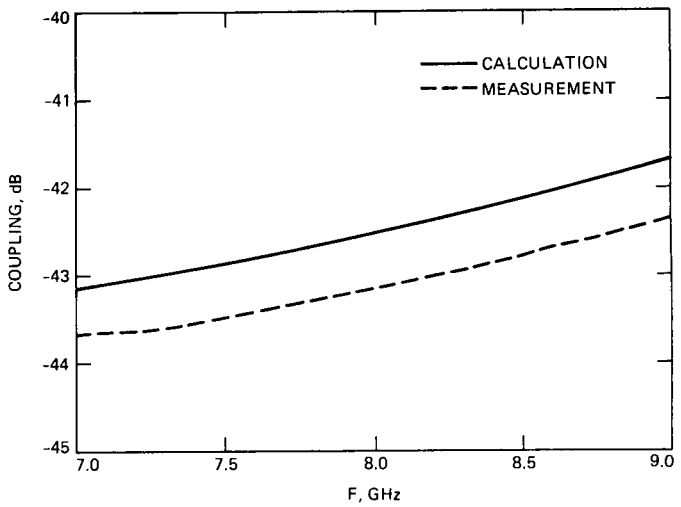


Fig. 3. Coupling of single-circular-aperture coupler.

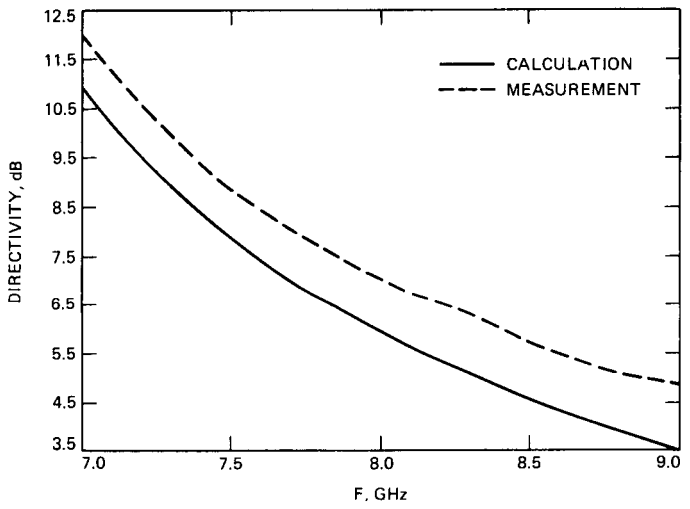


Fig. 4. Directivity of single-circular-aperture coupler.

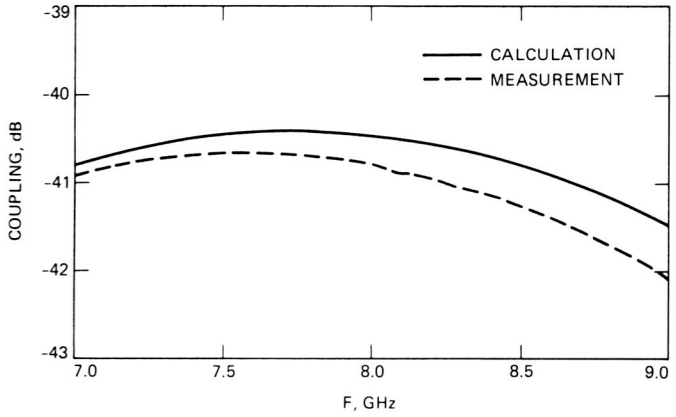


Fig. 5. Coupling of two-circular-aperture cross-guide coupler.

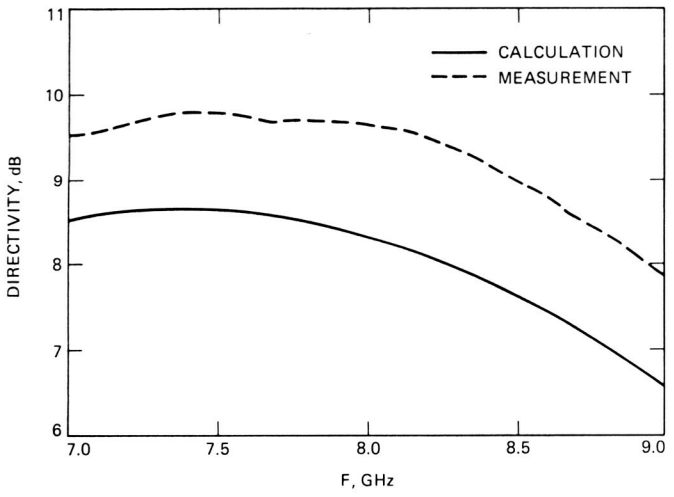


Fig. 6. Directivity of two-circular-aperture cross-guide coupler.

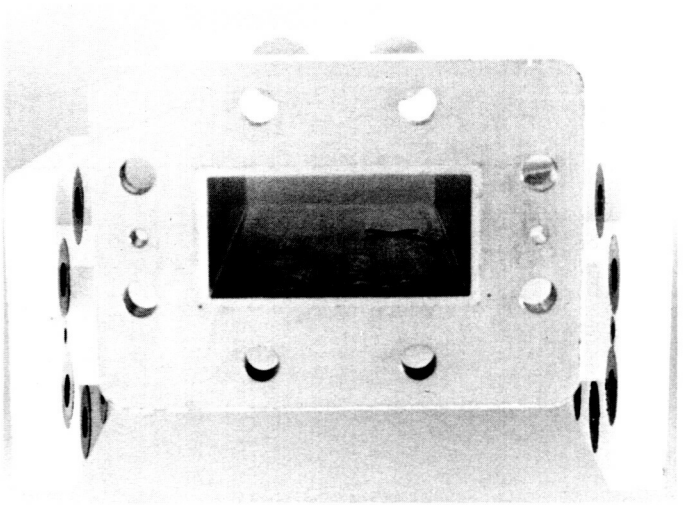


Fig. 7. A C-band two-cross-shaped-aperture cross-guide coupler.

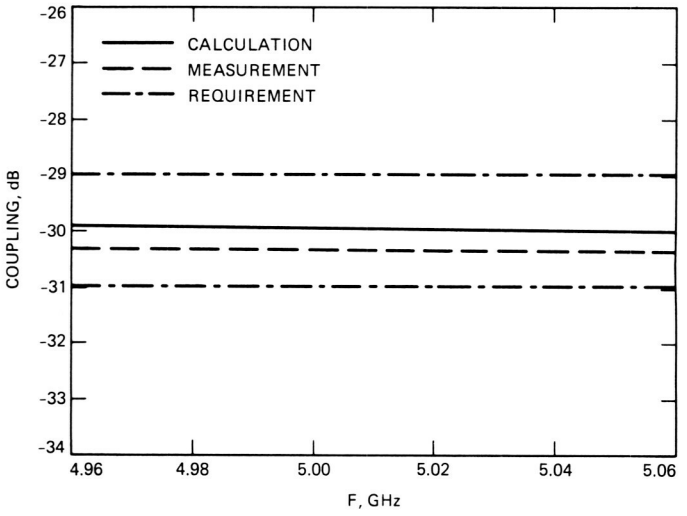


Fig. 8. Coupling of two-cross-shaped-aperture cross-guide coupler.

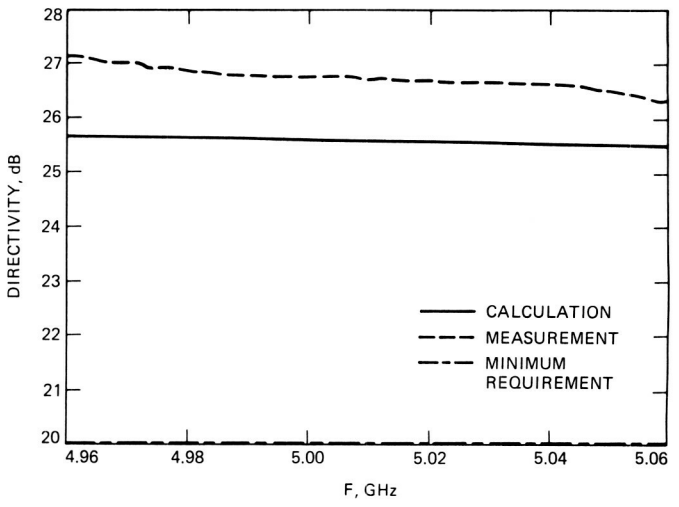


Fig. 9. Directivity of two-cross-shaped-aperture cross-guide coupler.

Appendix

Equations for Offset Aperture with an Arbitrary Angle θ

In this appendix, the results in [1] are extended to include an offset aperture with an arbitrary angle θ between two waveguides.

Referring to Figs. 1 and 2, assume the dimensions of the rectangular waveguides are a and b. The incident field in the primary waveguide is TE_{10} with amplitude A.

I. One-Offset Aperture with Arbitrary θ

The equivalent electric and magnetic dipoles $(\bar{P},\bar{M})$ located at the center of the aperture $(z = 0, x = d)$ for radiation into the secondary waveguide are

$$\bar{P} = \epsilon_0 p A \sin\frac{\pi d}{a}\hat{a}_y \tag{A-1}$$

$$\bar{M} = -m A Y_\omega\left(-\sin\frac{\pi d}{a}\hat{a}_x + j\frac{\pi}{\beta a}\cos\frac{\pi d}{a}\hat{a}_z\right) \tag{A-2}$$

where p and m are the electric and magnetic polarizability, Y_ω is the wave admittance for TE_{10} mode, and β is the propagation constant.

If the secondary waveguide with a coordinate system defined by unit vectors $\hat{a}'_x, \hat{a}'_y, \hat{a}'_z$ is rotated by an angle θ with respect to the primary waveguide with a coordinate system defined by unit vectors $\hat{a}_x, \hat{a}_y, \hat{a}_z$ (see Fig. 1)

$$\hat{a}_x = \cos\theta\,\hat{a}'_x + \sin\theta\,\hat{a}'_z \tag{A-3}$$

$$\hat{a}_y = \hat{a}'_y \tag{A-4}$$

$$\hat{a}_z = -\sin\theta\,\hat{a}'_x + \cos\theta\,\hat{a}'_z \tag{A-5}$$

Substituting Eqs. (A-3), (A-4), and (A-5) into Eq. (A-1) and (A-2)

$$\bar{P}' = \epsilon_0 p A \sin\frac{\pi d}{a}\hat{a}'_y$$

$$\bar{M}' = -m A Y_\omega\left[-\left(\sin\frac{\pi d}{a}\cos\theta + j\frac{\pi}{\beta a}\cos\frac{\pi d}{a}\sin\theta\right)\hat{a}'_x + \left(-\sin\frac{\pi d}{a}\sin\theta + j\frac{\pi}{\beta a}\cos\frac{\pi d}{a}\cos\theta\right)\hat{a}'_z\right]$$

The field radiated by the electric dipole has an amplitude

$$B_1(p,d,\theta) = \frac{j\omega}{abY_\omega}\bar{E}^{-'}_{10}\cdot\bar{p}'$$

$$= -\frac{j\omega\epsilon_0}{abY_\omega}pA\sin^2\frac{\pi d}{a}$$

$$= B_1(p,d)$$

in the coupled port and

$$B_2(p,d,\theta) = \frac{j\omega}{abY_\omega}\bar{E}^{+'}_{10}\cdot\bar{p}'$$

$$= B_1(p,d)$$

in the isolated port.

The field radiated by the magnetic dipole has amplitude

$$B_3(m,d,\theta) = \frac{j\omega}{abY_\omega}\bar{B}^{-'}_{10}\cdot\bar{M}'$$

$$= \frac{j\omega\mu_0 Y_\omega}{ab}mA\left[\left(\sin^2\frac{\pi d}{a} + \frac{\pi^2}{\beta^2 a^2}\cos^2\frac{\pi d}{a}\right)\cos\theta + 2j\frac{\pi}{\beta a}\sin\frac{\pi d}{a}\cos\frac{\pi d}{a}\sin\theta\right]$$

$$= B_3(m,d)\cos\theta + G(m,d)\sin\theta$$

in the coupled port and

$$B_4(m,d,\theta) = \frac{j\omega}{abY_\omega}\bar{B}^{+'}_{10}\cdot\bar{M}'$$

$$= \frac{j\omega\mu_0 Y_\omega}{ab}mA\left(-\sin^2\frac{\pi d}{a} + \frac{\pi^2}{\beta^2 a^2}\cos^2\frac{\pi d}{a}\right)\cos\theta$$

$$= B_4(m,d)\cos\theta$$

in the isolated port.

Therefore, the field in the secondary waveguide has an amplitude

$$B^{+}(p,m,d,\theta) = B_1(p,d) + B_3(m,d)\cos\theta + G(m,d)\sin\theta$$

in the coupled port and

$$B^{-}(p,m,d,\theta) = B_1(p,d) + B_4(m,d)\cos\theta$$

in the isolated port.

The coupling (C) and directivity (D) are given

$$C = 20\log\left(\frac{B^{+}(p,m,d,\theta)}{A}\right)$$

$$D = 20\log\left(\frac{B^{+}(p,m,d,\theta)}{B^{-}(p,m,d,\theta)}\right)$$

Some special cases are

(A) $d = \frac{a}{2}, \theta \neq 90^\circ$

$$G\left(m,\frac{a}{2}\right) = 0$$

$$B_3\left(m,\frac{a}{2}\right) = -B_4\left(m,\frac{a}{2}\right)$$

$$B^{+}\left(p,m,\frac{a}{2},\theta\right) = B_1\left(p,\frac{a}{2}\right) + B_3\left(m,\frac{a}{2}\right)\cos\theta$$

$$B^{-}\left(p,m,\frac{a}{2},\theta\right) = B_1\left(p,\frac{a}{2}\right) - B_3\left(m,\frac{a}{2}\right)\cos\theta$$

(B) $d = \frac{a}{2}, \theta = 90^\circ$

$$B^{+}\left(p,m,\frac{a}{2},90^\circ\right) = B^{-}\left(p,m,\frac{a}{2},90^\circ\right) = B_1\left(p,\frac{a}{2}\right)$$

$$D = 0\ \text{dB}$$

(C) $d \neq \frac{a}{2}, \theta = 0^\circ$

$$B^{+}(p,m,d,0^\circ) = B_1(p,d) + B_3(m,d)$$

$$B^{-}(p,m,d,0^\circ) = B_1(p,d) + B_4(m,d)$$

(D) $d \neq \frac{a}{2}, \theta = 90^\circ$

$$B^{+}(p,m,d,90^\circ) = B_1(p,d) + G(m,d)$$

$$B^{-}(p,m,d,90^\circ) = B_1(p,d)$$

II. Two Symmetrically Spaced Apertures with $\theta = 90$ Degrees

For two symmetrically spaced apertures which are located at $d_1 = d$ and $d_2 = a - d$, let $\Delta = a - 2d$, $\theta = 90^\circ$ (Fig. 2). The field radiated in the secondary waveguide is

$$BB^{+}(p,m,d,90^\circ) = B^{+}(p,m,d,90^\circ) + B^{+}(p,m,a-d,90^\circ)e^{-2j\beta\Delta}$$

in the coupled port and

$$BB^{-}(p,m,d,90^\circ) = [B^{-}(p,m,d,90^\circ) + B^{-}(p,m,a-d,90^\circ)]\,e^{-j\beta\Delta}$$

in the isolated port.

Modal Analysis Applied to Circular, Rectangular, and Coaxial Waveguides

D. J. Hoppe
Radio Frequency and Microwave Subsystems Section

This report summarizes recent developments in the analysis of various waveguide components and feedhorns using Modal Analysis (Mode Matching Method). A brief description of the theory is presented, and the important features of the method are pointed out. Specific examples in circular, rectangular, and coaxial waveguides are included, with comparisons between the theory and experimental measurements. Extensions to the methods are described.

I. Introduction

Modal Analysis has been shown to be a highly accurate and versatile method for analyzing a wide variety of waveguide devices [1]-[4]. The method is capable of accounting for multiple reflections within the device, stored energy at each discontinuity, and higher-order mode propagation if it occurs. Its high accuracy makes it useful for tolerance studies after a final design has been determined. This report compares computed and experimental results for the scattering parameters of three examples. One example is taken from each of the following waveguide types: rectangular waveguide, circular waveguide, and coaxial waveguide propagating the TE_{11} mode.

II. Theory

The theory described below is well known and is summarized in [1]. In applying the modal analysis method, the waveguide device is broken up into a series of sections that are joined by a step discontinuity as is shown in Fig. 1. For smooth changes in waveguide dimensions, the change is approximated by a large number of steps. At this point, the type of waveguide is arbitrary but the common area between the two guides must be identical to the cross-section of the smaller waveguide. This eliminates a class of offset connections but is usually not important for analyzing a practical device. In addition, for the circular waveguide and the coax, all guides are required to possess the same center line. This simplifies the analysis since only modes with one azimuthal variation need to be considered. Again, this is not restrictive for most practical applications. Next, the development of some of the important equations is presented. In Fig. 1, the fields to the left of the junction ($z < 0$) are represented as a sum of the normal modes of waveguide I.

$$\underline{E}_I = \sum_{m=1}^{M} (A_{Im}\, e^{-j\beta_m z} + B_{Im}\, e^{j\beta_m z})\underline{e}_{Im} \qquad (1)$$

$$\underline{H}_I = \sum_{m=1}^{M} (A_{Im}\, e^{-j\beta_m z} - B_{Im}\, e^{j\beta_m z})\underline{h}_{Im} \qquad (2)$$

Here M is chosen large enough for convergence, and $\underline{e}_{Im}$ and $\underline{h}_{Im}$ are the normalized vector functions for the mth mode. For example, in a circular waveguide, $m = 1 = TE_{11}$, $m = 2 = TM_{11}$, $m = 3 = TE_{12}$, etc. A_{Im} represents the magnitude of the forward traveling mth mode, and B_{Im} the magnitude of the reverse traveling mth mode.

The normalization of $\underline{e}_{Im}$ and $\underline{h}_{Im}$ is such that

$$\int_{S_I} (\underline{e}_{Im} \times \underline{h}_{Im}) \cdot ds = R_{mm} \tag{3}$$

and from the orthogonality of the waveguide mod

$$\int_{S_I} (\underline{e}_{Im} \times \underline{h}_{In}) \cdot ds = 0 \qquad m \neq n \tag{4}$$

Similarly, in region II

$$\underline{E}_{II} = \sum_{n=1}^{N} (A_{IIn}\, e^{j\beta_n z} + B_{IIn}\, e^{-j\beta_n z})\, \underline{e}_{IIn} \tag{5}$$

$$\underline{H}_{II} = \sum_{n=1}^{N} -(A_{IIn}\, e^{j\beta_n z} - B_{IIn} e^{-j\beta_n z})\, \underline{h}_{IIn} \tag{6}$$

where N is the number of modes chosen in region II. Relations analogous to Eqs. (3) and (4) hold for this region also.

Matching the electric and magnetic fields over the common aperture results in the following scattering matrix equation.

$$B = [S]\, A \tag{7}$$

$$B = \left[\frac{B_I}{B_{II}}\right] \tag{8}$$

$$A = \left[\frac{A_I}{A_{II}}\right] \tag{9}$$

and

$$[S] = \begin{bmatrix} [S_{11}] & [S_{12}] \\ [S_{21}] & [S_{22}] \end{bmatrix} \tag{10}$$

In Eqs. (8) and (9), B_I and B_{II} are vectors containing the reflected-mode amplitudes, while A_I and A_{II} contain the incident-mode amplitudes. The derivation of these equations is given in the Appendix.

For the normalized vectors, the power carried by the mth forward traveling mode is given by $|A_{Im}|^2$, and for the mth reverse traveling mode is $|B_{Im}|^2$.

Next, the matrices for a straight section of length L are needed. The solution is trivial, giving

$$[S_{11}] = [0] \tag{11}$$

$$[S_{12}] = [\gamma_{12}] \tag{12}$$

$$[S_{21}] = [\gamma_{21}] \tag{13}$$

$$[S_{22}] = [0] \tag{14}$$

where $[0]$ is the zero matrix and $[\gamma_{12}]$ and $[\gamma_{21}]$ are diagonal matrices with elements

$$\gamma_{21}(n,n) = \bar{e}^{\gamma_n L} = \gamma_{12}(n,n) \tag{15}$$

γ_n being the propagation constant for the nth mode and L being the section length.

This completes the summary of the required equations for each step in the analysis. Using these results, matrices for each step and each straight section in the device are determined and then combined using equations in [1]. At the completion of the analysis, the overall matrix is obtained, relating the normalized output vectors B_I and B_{II} at the ends of the device to the normalized input vectors A_I and A_{II}.

$$B_I = [S_{11}]\, A_I + [S_{12}]\, A_{II} \tag{16}$$

$$B_{II} = [S_{21}]\, A_I + [S_{22}]\, A_{II} \tag{17}$$

In many situations, a set of modes is incident only on the left end of a device and one wants to determine the reflected and transmitted modes. The user specifies the input mode vector A_I, $A_{II} = 0$, and Eqs. (16) and (17) become

$$B_I = [S_{11}]\, A_I \tag{18}$$

$$B_{II} = [S_{21}]\, A_I \tag{19}$$

III. Results

Computer programs [5] have been written to carry out the above calculations for rectangular, circular, and coaxial waveguides. For the junctions involved, the integrals of Eq. (A-6) can be carried out in closed form, which greatly simplifies the programming. Three examples, one in each waveguide type, are used to demonstrate the excellent agreement between theory and experiment. All of the measurements were made using an HP 8510 network analyzer and a full two-port calibration.

The first example, shown in Fig. 2, is a circular waveguide transition. For the theoretical results, 25 modes were used in the input section. The number of modes used in subsequent sections is chosen by the program for optimum convergence [1]. The return loss measurements (Fig. 3a) are typically within 0.2 dB, except near the minimum reflection point at 8.25 GHz. Phase results (Fig. 3b) are also in close agreement, typically within a few degrees across the band. For nearly every observation point, the difference between theory and measurement is within the accuracy specification of the network analyzer. Slight inaccuracies in the waveguide dimensions and rounding of some of the corners can also account for the small disagreement that remains. Figure 4 illustrates the convergence of the solution at 8 GHz as the number of modes used in the input section is increased. From these plots, we see that the solution has stabilized once 20 or more modes are used in the input waveguide. The number of modes required for convergence depends on the particular device, but in general larger waveguides with respect to a wavelength and thin irises require that more modes be used in order to get the same accuracy.

A rectangular waveguide example is shown in Fig. 5 and both theoretical and experimental results are given in Fig. 6. The device is a WR125 to 0.8-inch square-to-WR125 transition that was fabricated for use in the ring resonator at 8.51 GHz. The device consists of nine waveguide sections. The figure shows that, as in the previous example, the theoretical and experimental results are in excellent agreement. Only slight discrepancies appear near the minimum reflection point. More modes may be needed to represent the field in this region, or else the 0.030-inch radius on all corners, which was not accounted for in the calculation, may have a stronger effect in this frequency band. For this example, modes with first index m less than or equal to 7 and second index n less than or equal to 6 were used in the input guide. As with the circular waveguide program, maximum mode indices in the following sections are chosen according to waveguide size and symmetry considerations.

The final example is the coaxial iris shown in Fig. 7, with theoretical and experimental return loss results shown in Fig. 8. The coaxial region is excited by a TE_{11} circular waveguide mode that excites only the higher-order coax modes with first index equal to 1; the normal TEM coax mode is not excited in this case. Measurements of the iris were made by calibrating in the circular waveguide and using the time domain gating features of the HP 8510 network analyzer to isolate the reflections from only the iris. The only other complication associated with the coax is that a transcendental equation must be solved for each mode in each section in order to determine a cutoff wavelength. This increases the computation time required to solve a coax problem compared to a similar circular or rectangular waveguide problem. As with the previous examples, the agreement between theory and experiment is good, particularly considering the errors introduced by using the time domain features of the HP 8510.

IV. Conclusion

Three representative examples have been given to demonstrate the accuracy of the modal analysis method. A large number of waveguide devices such as horns, corrugated waveguides, transitions, filters, and smooth tapers can be analyzed using these programs. In addition, several extensions have been made to the codes in order to allow for differing dielectric constants in the sections, making them useful for window design. For large smooth-wall or corrugated horns, the reflection at the aperture may be neglected, and the far-field pattern can be found from the propagating modes in the aperture. In addition, the important case of ring-loaded slots [6], which is a combination of the coaxial and circular program, has also been programmed, but no experimental results are presently available.

Acknowledgments

The author would like to acknowledge the assistance of Dr. Farzin Manshadi in the development of the rectangular waveguide code, and Phil Stanton for providing the experimental data for the coaxial waveguide example.

References

[1] G. L. James, "Analysis and Design of TE_{11} to HE_{11} Corrugated Cylindrical Waveguide Mode Converters," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-29, pp. 1059–1066, October 1981.

[2] E. Huhn and V. Hombach, "Computer-Aided Analysis of Corrugated Horns with Axial or Ring-Loaded Slots," *IEE Conf. Publ. 219 (ICAP 83) Part 1*, pp. 127–131, 1983.

[3] G. L. James, "Admittance of Irises in Coaxial and Circular Waveguides for TE_{11}-Mode Excitation," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-35, pp. 430–434, April 1987.

[4] H. Patzelt and F. Arndt, "Double-Plane Steps in Rectangular Waveguides and Their Application for Transformers, Irises, and Filters," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-30, pp. 771–776, May 1982.

[5] D. Hoppe, "Scattering Matrix Program for Circular Waveguide Junctions," in *Cosmic Software Catalog*, 1987 edition, NASA-CR-179669, NTO-17245, Georgia: NASA's Computer Software Management and Information Center, 1987.

[6] G. L. James and B. M. Thomas, "TE_{11} to HE_{11} Cylindrical Waveguide Mode Converters Using Ring-Loaded Slots," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-30, pp. 278–285, March 1982.

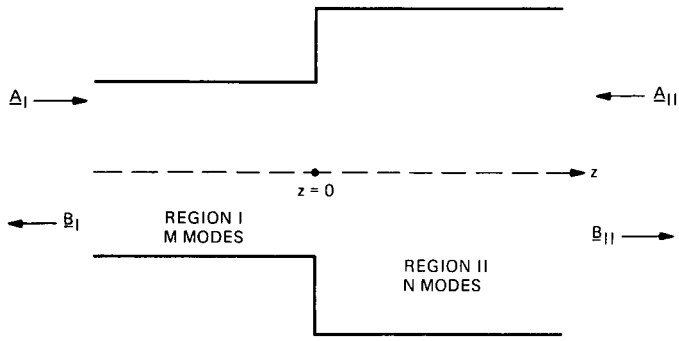


Fig. 1. Parameters for a single junction.

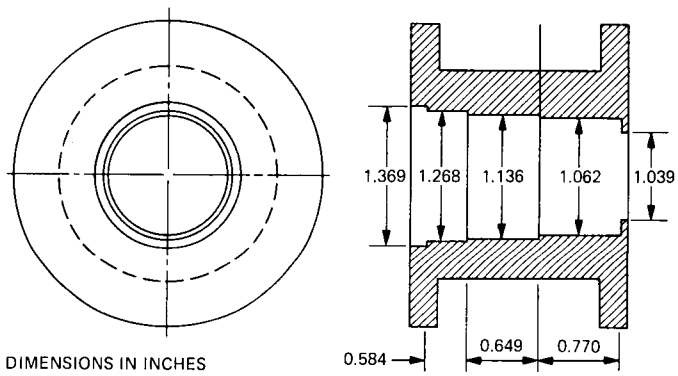


Fig. 2. Circular waveguide example.

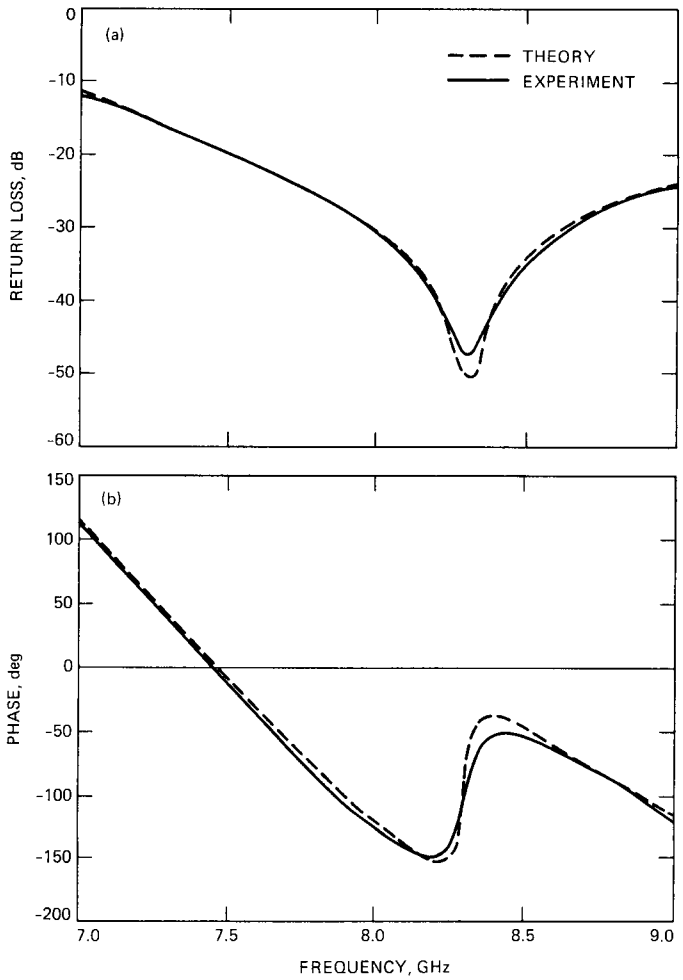


Fig. 3. Circular waveguide results: (a) return loss and (b) phase.

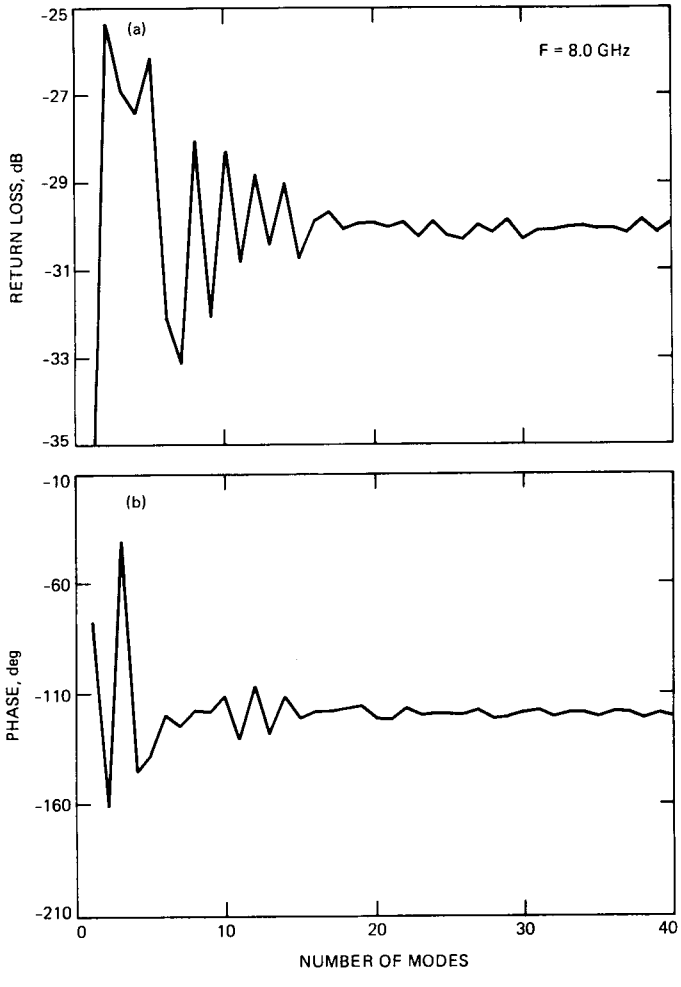


Fig. 4. Convergence for circular waveguide example: (a) return loss and (b) phase.

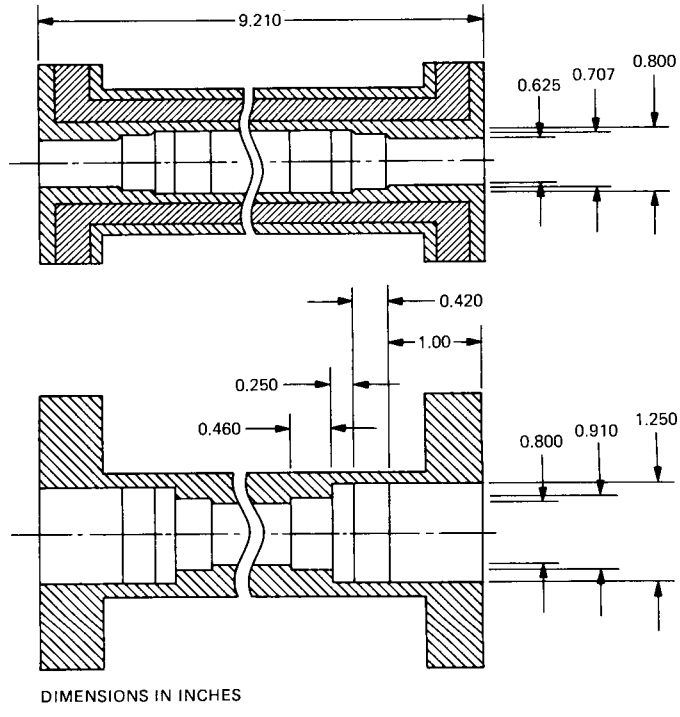


Fig. 5. Rectangular waveguide example.

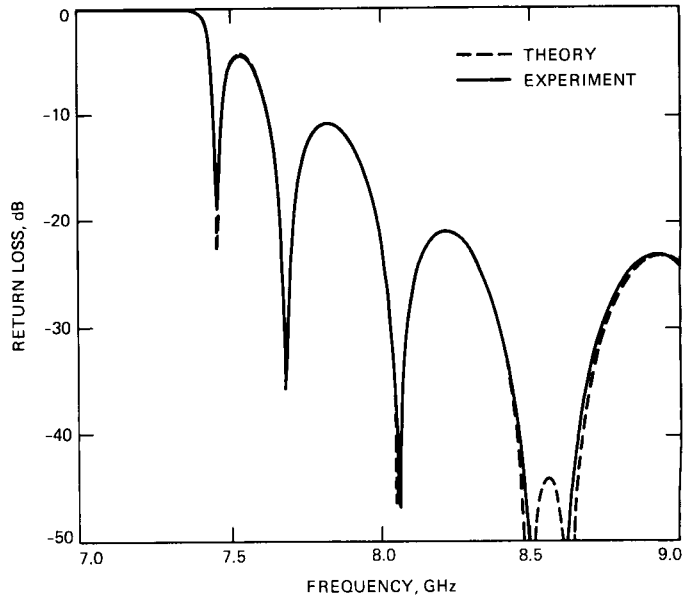


Fig. 6. Rectangular waveguide return loss results.

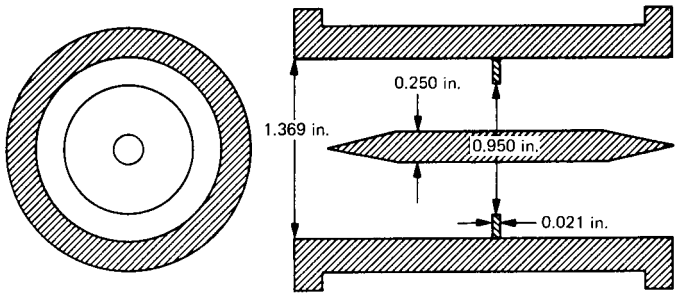


Fig. 7. Coaxial waveguide example.

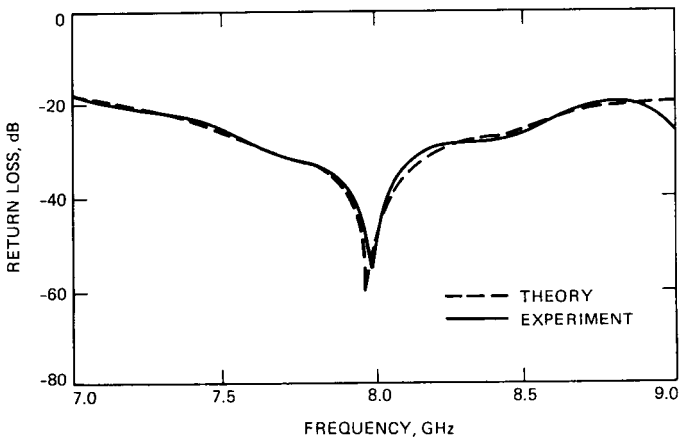


Fig. 8. Coaxial waveguide return loss results.

Appendix

Derivation of the Waveguide Scattering Matrix Equation

To derive Eq. (10), the electric fields inside the common aperture between the two regions are matched.

$$\underline{E}_I = \underline{E}_{II} \text{ inside } S_I \tag{A-1}$$

$$\underline{E}_I \times \underline{h}_{IIn} = \underline{E}_{II} \times \underline{h}_{IIn} \text{ inside } S_I \tag{A-2}$$

$$\int_{S_I} (\underline{E}_I \times \underline{h}_{IIn}) \cdot ds = \int_{S_I} (\underline{E}_{II} \times \underline{h}_{IIn}) \cdot ds \tag{A-3}$$

Since $E = 0$ on the conductor making up the surface $S_{II} - S_I$, the integral on the right-hand side of Eq. (A-3) may be extended over S_{II}.

$$\int_{S_I} (\underline{E}_I \times \underline{h}_{IIn}) \cdot ds = \int_{S_{II}} (\underline{E}_{II} \times \underline{h}_{IIn}) \cdot ds \tag{A-4}$$

Using the properties in Eqs. (1)–(6) the following is obtained:

$$\sum_{m=1}^{M} (A_{Im} + B_{Im}) P_{mn} = (A_{IIn} + B_{IIn}) Q_{nn} \tag{A-5}$$

where

$$P_{mn} = \int_{S_I} (\underline{e}_{Im} \times \underline{h}_{IIn}) \cdot ds \tag{A-6}$$

and

$$Q_{nn} = \int_{S_{II}} (\underline{e}_{IIn} \times \underline{h}_{IIn}) \cdot ds \tag{A-7}$$

The other boundary condition needed is

$$\underline{H}_I = \underline{H}_{II} \text{ within } S_I \tag{A-8}$$

Following a similar line of reasoning

$$\int_{S_I} (\underline{e}_{Im} \times \underline{H}_I) \cdot ds = \int_{S_I} (\underline{e}_{Im} \times \underline{H}_{II}) \cdot ds \tag{A-9}$$

giving

$$R_{mm}(A_{Im} - B_{Im}) = \sum_{n=1}^{N} P_{mn}(B_{IIn} - A_{IIn}) \tag{A-10}$$

$$R_{mm} = \int_{S_I} (\underline{e}_{Im} \times \underline{h}_{Im}) \cdot ds$$

Equations (A-5) and (A-10) may be recast into a more compact matrix form, giving

$$[P]\,(A_I + B_I) = [Q]\,(A_{II} + B_{II}) \tag{A-11}$$

$$[R]\,(A_I - B_I) = [P]^T\,(B_{II} - A_{II}) \tag{A-12}$$

Here $[P]^T$ is the transpose of the matrix $[P]$, and $[R]$ is an $m \times m$ diagonal matrix, and $[Q]$ is an $n \times n$ diagonal matrix.

Next, Eq. (A-12) is converted into a scattering matrix format relating the normalized output vectors B_I and B_{II} to the normalized input vectors A_I and A_{II}.

The submatrices $[S_{11}]$, $[S_{12}]$, $[S_{21}]$, and $[S_{22}]$ are derived from the $[P]$, $[P]^T$ $[R]$, and $[Q]$ matrices by simple matrix math and Eqs. (A-11) and (A-12).

$$[S_{11}] = [\sqrt{R}]\,([R] + [P]^T\,[P])^{-1}\,([R] - [P]^T\,[P])\,[\sqrt{R}]^{-1} \tag{A-13}$$

$$[S_{12}] = 2[\sqrt{R}]\,([R] + [P]^T\,[P])\,[P]^T\,[\sqrt{Q}]^{-1} \tag{A-14}$$

$$[S_{21}] = 2[\sqrt{Q}]\,([Q] + [P]\,[P]^T)\,[P]\,[\sqrt{R}]^{-1} \tag{A-15}$$

$$[S_{22}] = [Q]\,([Q] + [P]\,[P]^T)^{-1}\,([Q] - [P]\,[P]^T)\,[\sqrt{Q}]^{-1} \tag{A-16}$$

In these equations, $[I]$ represents the unit matrix and $[\sqrt{R}]\,[\sqrt{R}] = [R]$, and $[\sqrt{Q}]\,[\sqrt{Q}] = [\sqrt{Q}]$ are from Eqs. (3) and (7). These factors form the normalization of the vectors A and B. This completes the solution for the junction between the two different waveguides.

Conceptual Design of a 1-MW CW X-Band Transmitter for Planetary Radar

A. M. Bhanji, D. J. Hoppe, B. L. Conroy, and A. J. Freiley
Radio Frequency and Microwave Subsystems Section

A proposed conceptual design to increase the output power of an existing X-band radar transmitter used for planetary radar exploration from 365 kW to 1 MW CW is presented. The paper covers the basic transmitter system requirements as dictated by the specifications for the radar. The characteristics and expected performance of the high-power klystron are considered, and the transmitter power amplifier system is described. Also included in the discussion is the design of all of the associated high-power microwave components, the feed system, and the phase-stable exciter. The expected performance of the beam supply, heat exchanger, and monitor and control devices is also presented. Finally, an assessment of the state-of-the-art technology needed to meet system requirements is given and possible areas of difficulty are summarized.

I. Introduction

Radar has been used as a remote tool for exploration of our solar system since 1946, when detection of echoes from the Moon was reported. Since then, ground-based radar studies have been made of the planets Mercury, Venus, and Mars, the Galilean satellites, the rings of Saturn, and nearly a dozen asteroids. The sensitivity of the radar instruments used in these experiments has increased by a factor of approximately 10^{12} since the first lunar detection. Such great gains in sensitivity have been achieved by extraordinary improvements in antenna size, low-noise receiving systems, high-speed digital signal processing, and higher-power transmitters at higher frequencies.

One area that needs further development in order to increase the remote sensing capability of radar is the increase of power in the transmitter. Therefore, at JPL, design and development has been initiated to extend the present X-band transmitter capability from 365 kW CW to 1 MW CW.

This experimental transmitter will be installed on the 70-meter dual-reflector antenna at Goldstone, California, which is equipped with a rotatable, asymmetric subreflector that permits the use of multiple feed systems at the antenna focus. The subreflector can be precision indexed to a fixed number of positions that will allow each feed to properly illuminate the main reflector. Assuming an aperture efficiency of about

75% and corresponding antenna gain of 74.5 dB at X-band, with the proposed 1-MW CW transmitter, the X-band radar system will have an effective radiated power of about 28 TW (2.8×10^{13} W).

Using an assessment of the state-of-the-art technology, this article describes the X-band transmitter, including the transmission system and a feed system.

II. Transmitter System Requirements

The transmitters used for radar astronomy systems differ from conventional radar systems in that they require high average power, rather than high peak power, over the bandwidth required to handle the transmitted signal [1]. It is also important that these transmitters be coherent in order to determine the phase relationships of the returned signals, and they must have high phase stability if coherent measurements are to be made over long periods of time. The transmitter must also be capable of modulation by a variety of pulse programs, while maintaining the phase and amplitude fidelity and pulse-to-pulse stability required for pulse-compression systems incorporated in the radar.

The above requirements illustrate that high power alone will not provide the desired CW radar transmitter capabilities. If this were the case, it might be more easily obtained with an oscillator rather than an amplifier. Besides the advantage of having dynamic control of amplitude and phase, the appeal of using an amplifier is that it eliminates the need for phase-locking an oscillator to a control signal.

Based on the above requirements, the X-band radar transmitter specifications are given in Table 1.

III. The Transmitter System

As shown in Fig. 1, the transmitter will include a power supply that converts 2400-V, 3-phase, 60-Hz line voltage to direct current at up to 50 kV with a power limitation of 2.25 MW for the four klystron amplifier beams. The frequency synthesizer-based exciter and the buffer amplifier will provide an input signal to these klystrons, and each of the four klystrons will provide approximately a 53-dB power gain. The automated transmitter control will perform monitoring and control of all functions. Protective devices (interlocks) will prevent damage to equipment by removing voltage and in some cases drive power in the event of a malfunction. The liquid-to-air 2.5-MW heat exchanger will be used to cool the amplifier, the power supply, various auxiliaries to the transmitter, microwave components of the transmitter, and microwave components of the transmission line. The following paragraphs describe each of the above components in greater detail.

A. Exciter

Figure 2 shows the proposed configuration of the exciter portion of the 1-MW radar, which is based on an HP 8662A synthesizer. The synthesizer uses the 10-MHz reference signal from a hydrogen maser to produce a phase-coherent output at 640 MHz, and a phase-coherent variable-frequency signal from 10 kHz to 640 MHz with 0.1-Hz resolution, or from 640 to 1280 MHz with 0.2-Hz resolution. The 640-MHz signal can be picked off and mixed with the variable-frequency output. For the transmitter, the variable-frequency output is set to 830 MHz and mixed with 12 times the 640 MHz to produce 8510 MHz. In the receive configuration, the variable frequency is set to 505 MHz and mixed with 12 times the 640 MHz to produce a signal at 8185 MHz, which could be used as the first local oscillator in the receiver. A similar system, using 3 times 640 MHz, is used for the S-band exciter and could also be used as a first local oscillator for an S-band receiver.

Reduction of phase noise is a major concern in the exciter design. By mixing the output of an extremely low-noise, high-frequency synthesizer with low multiples of a clean, fixed-frequency oscillator, total phase noise should be reduced by more than 20 dB from that produced by the more standard method of using a high multiple of a low-frequency synthesizer.

Provisions are made for biphase noise modulation, frequency hopping, and continuous frequency sweep for Doppler cancellation. The frequency hopping is accomplished by sending frequency step commands to the synthesizer on an IEEE-488 interface. Doppler cancellation is done through a combination of a coarse frequency, set through the IEEE-488 interface, and an analog voltage at the search oscillator input. Phase noise modulation is done directly with a biphase modulator at the output frequency.

A power divider is included in the exciter, providing separate outputs for the phase reference system and the klystron drive system in the buffer amplifier.

B. Buffer Amplifier

The buffer amplifier is the second major block of Fig. 2. For the first choice of transmission line arrangements (see Section D), phase-tracking loops and electronic polarization control are provided in the exciter. The phase-tracking loop uses a voltage-tracking controlled phase shifter in the drive path of each klystron to compensate for phase changes in klystrons or microwave components. A sample of the output from each

klystron is taken as close to the feedhorn as possible and compared with the reference signal. Symmetry in the waveguide paths from the final splitter is still required to prevent differential phase shift between the two horn inputs driven by the same klystron.

Because the feed system uses two klystrons to drive the in-line inputs to the horn and the other two klystrons to drive the orthomode inputs, polarization control is achieved by the phase shifter after the first splitter in the buffer amplifier. Phase shifts of ±90 degrees yield right-hand or left-hand circular polarization, and phase shifts of 0 or 180 degrees yield orthogonal linear polarizations. Other phase shift values produce various elliptical polarizations. A corresponding phase shift must be introduced in the phase correction loop, but in the case of switching from right-hand circular to left-hand circular polarization, this can be done with an inverting amplifier after the phase detector. For the waveguide-based system, no special electronic control is required in the exciter. In both cases, a solid-state amplifier for each klystron provides the approximately 2-watt drive power required.

C. Power Amplifier

The power amplifier section of the transmitter contains four 250-kW CW klystrons. The requirements of high power, high gain, good efficiency, ease of modulation, and an output spectrum free from spurious signals and noise make a klystron linear-beam tube the natural choice for radars, as long as its narrow bandwidth, high operating voltage, and large size can be tolerated.

Early in 1986, Varian Associates was contracted by JPL to do a paper design for eventual development and production of extremely high-power X-band klystrons. The characteristics of this tube, designated VKX-7864A, are given in Table 2.

As part of the design for this klystron, Varian is expected to provide phase modulation sensitivity due to various pushing factors such as beam voltage, drive power, body coolant, focus current, and heater voltage. These phase-pushing factors are important because they produce unwanted discrete lines on the phase noise spectrum.

Based on overall transmitter system specifications, these modulation sensitivities, together with modulation sensitivity of other pushing factors, establish the requirements for power supply stability, body coolant temperature stability and stabilities of other pushing (control) factors.

Each klystron is provided with an arc detector at the window and a reverse power coupler for protection. In the event of a fault, these monitors will remove the RF drive before permanent damage can occur.

One of the critical elements of the klystron is the guiding magnet. This device is a solenoid which surrounds the interaction volume and keeps the electron beam focused in the tube length before the collector. A control of better than 1% must be exercised to maintain high efficiency and low body current. This solenoid will weigh about 300 pounds and will require a 300-V, 20-A dc power supply to provide the proper magnetic field.

The separate coolant manifold for each klystron will monitor and control flows, temperature, and pressure. This data will be routed to the data collection system, which is described in Section H.

The power amplifier, including the transmission line (described in the next section) will all be housed in the cone. The mechanical layout is shown in Fig. 3.

D. Transmission Line

Two alternatives for the transmission line system are under consideration. The first system is shown in Fig. 4. For this system, the output from each klystron passes through a waveguide switch and a directional coupler before being split into two 125-kW signals. Four of these signals (two pairs) feed the in-line ports of four orthomode junctions, while the other pair feed the ortho ports. Thus, by adjusting the relative phase between the two pairs of klystrons, one of two orthogonal linear polarizations, RHCP or LHCP, may be obtained (see buffer amplifier, Section C). The outputs of the four orthomode junctions then feed the four inputs to the multimode feedhorn. In this system, phase detectors and electrically controlled phase shifters will be used to adjust the outputs of each of the klystrons and to provide polarization control. The reliance on electronics reduces the complexity of the waveguide layout in comparison to the waveguide-based system described below.

Figure 5 shows the waveguide-based alternative to the previous system. This system is similar to that used at the Haystack Hill Observatory in Westford, Mass. [2]. In this system, a series of splitters and combiners ultimately forms four identical signals. Each of the signals is composed of 25% of the power from each klystron, and all four are equal in amplitude and phase, independent of the four klystron outputs. Phase shifters are used in each of the drive lines to the klystrons in order to minimize the power in the waster loads. This adjustment is made once, and if klystron parameters or frequencies change during a track, only the waster-load power will change.

Polarization is changed mechanically through switches immediately before the orthomode junctions. Although this approach is complicated mechanically, it has the advantage that beam position and polarization purity are virtually guaranteed without the use of any electronics.

The final decision on which of the two possible systems, or a combination thereof, will be used depends on how closely the four klystron tubes can be matched in terms of phase, gain, and group delay versus frequency. Measurements on the two existing tubes as well as discussions with the tube manufacturer (Varian Assoc.) are underway to help answer these questions. The effects of aging must also be considered to guarantee that the system will run reliably over the expected lifetime of the tubes with minimal adjustment.

In both systems, WR125 is used as the high-power waveguide to avoid operation close to the higher-order modes in WR137, which begin propagating at 8600 MHz [3].

The electrically operated waveguide switches allow selection of the radar antenna or water loads for termination of transmitter output power. The water loads will also be used for calibrating the output power calorimetrically.

On receive, the klystrons are turned off and the switches are rotated, connecting the receive waveguide to the feedhorn. Through a series of combiners, RHCP and LHCP signals are formed. These signals then enter the dual-channel maser, and finally the heterodyne receiver.

In addition, the existing low-noise system, which uses a separate corrugated horn for receiving, will be retained. The disadvantage of this receiving arrangement is that the antenna subreflector must be rotated between transmit and receive cycles.

E. Antenna Feed System

The final component required in the transmission line for the radar system is the feedhorn, which will launch the transmitter power toward the subreflector of the antenna. The horn should illuminate the subreflector efficiently and increase the overall noise temperature of the system as slightly as possible. The horn will be designed to meet or exceed the RF performance of the feedhorns presently in use, with the added feature of 1-MW capability.

Several possible feed types were considered for the 1-MW system. Experience indicates that conventional corrugated or dual-mode horns are not capable of handling the large CW power without breaking down at the small-diameter input section. It was also found that a rather large number of small horns arranged in a closely packed array would be required to illuminate the subreflector efficiently. Due to the complexity of this type of system, as well as the losses associated with the power-splitting components, the array concept was also rejected for the 1-MW system.

The chosen design uses a multiflare rectangular horn [4]. Such a horn is well suited to the 1-MW system, since it possesses an excellent radiation pattern and has been used in other power applications.

The multimode horn is depicted in Fig. 6. Four square waveguides feed a large square chamber where the power is launched into a square multiflare horn. Since the large chamber is oversized for the frequency of interest, the sum of the power in the four waveguides can be supported without breakdown. In the present case, each of the guides must support 250 kW, and the large chamber 1 MW. Flare angle changes are used to generate the required higher-order modes for pattern symmetry.

The analysis of the horn is carried out using mode-matching methods. The overall scattering matrix of the horn is obtained and, using these results, the input match, as well as the far-field radiation pattern, can be predicted for arbitrary input levels and phases in the four input guides. Calculated radiation patterns for the horn at 8.51 GHz are shown in Fig. 7. A detailed description of the horn will be provided in a future report.

Estimates for the peak electric fields in the horn indicate that the maximum level (about 6.9 kV/cm) occurs in the four feeding waveguides, which are 0.95 in. square. This should be compared to the present 400-kW WR125 waveguide system (about 8.5 kV/cm) and the theoretical limit for a 0.95-in.-square waveguide, which is about 2.1 MW. Resonant ring tests [5] will be used to evaluate the power performance of the orthomode junctions and the horn. Should arcing become a problem, backup approaches include evacuating areas of the horn or pressurizing them with a dielectric gas such as SF_6.

F. Beam Power Supply

A block diagram of the beam supply is shown in Fig. 8. Power at 12,600 V, 3 phase, and 60 Hz is supplied to separate substations from the commercial line (Edison), which is underground for the last mile. The 2400-V substation supplies the main motor generator only. Under critical operation, this 2400-V, 60-Hz power is supplied separately by the diesel generator. All auxiliaries are supplied from a 480-V substation. The output of the main motor-generator at 400 Hz is stepped up in voltage, transformed, rectified, and delivered to the klystron load through a filter, crowbar, and series limiter

resistor to voltages of 50 kV and 2.25 MW. The design of this power supply will maintain output ripple under full load of less than 0.05%, regulation of 0.01%, and settling time of 200 milliseconds.

The use of a frequency converter (such as a motor-generator) might seem unnecessary, but it actually provides worthwhile technical and economic advantages. It isolates the power line from a crowbar of the dc supply and greatly simplifies line protection problems. It also isolates the supply from short-duration line voltage fluctuations and transients due to the large inertia of the rotating components. The change from 60 to 400 Hz reduces all transformer and filter sizes and costs.

This beam supply is required to provide 50 kV between the klystron collector and cathode at a beam current of 11 amps (per klystron) during the long radar pulses (up to 10 hours).

The ability of the beam supply to remain ripple free and tightly regulated, and to settle in 200 msec will require a unique and state-of-the-art feedback control circuit. The design is in progress and the first approach is being tested. The supply must also be capable of withstanding the stress imposed on it when an arc occurs in any klystron.

G. Cooling System

The cooling system provides a 2.5-MW cooling capacity for klystrons, focusing magnets, high-power microwave components, water loads, feed, and the transformer rectifier, including the motor-generator clutch. Basically, the cooling system is a closed loop which consists of a heat exchanger, a distribution manifold, and all connecting piping. Coolant is circulated through the cooling system by the heat exchanger pumps. The coolant gains heat as it passes through the RF system (buffer amplifier, power amplifier, and microwave components) and loses heat as it passes through the heat exchanger. A purity loop is connected to the input of the heat exchanger to maintain the purity of the coolant.

As part of this transmitter modification, the pumps will have to be upgraded, including replacing all of the 6-inch water lines, and a complex water-switching mechanism will have to be installed. The water-to-water heat exchanger will have to be changed to a liquid-to-air heat exchanger.

H. Monitor and Control

Operation of the 1-MW radar will be extensively automated. The control system will be composed of an HP Industrial Vectra computer, two HP 38526 Data Acquisition and Control units, a frequency counter, and a multichannel power meter. All the instruments will be connected by an IEEE-488 interface bus, with a fiber optic extension between the control room and the cone area of the antenna. The IEEE bus will also connect to the synthesizer in the exciter.

The monitor and control software will be written in Ada, a language especially designed for hardware control applications, and will make use of artificial intelligence principles to maximize the system functionality while maintaining a simple user interface. It will automatically keep long-term data logs to look for trends and predict failures before they can disable the radar. It will correlate data from different sources to distinguish between sensor problems and transmitter problems, and it will be able to calibrate the RF power meters by precision measurement of the flow rates and temperature rises of the coolant in the water loads.

IV. Concerns and Conclusion

The previous sections make it clear that this radar transmitter will be a very complicated system, and therefore there are several areas of concern. Of primary concern is the possibility of waveguide/feedhorn breakdown. Associated with this problem is the difficulty of testing the components at or above the normal operating power level before all of the klystrons arrive. Resonant ring testing can be used for some components, but the most critical component, the feedhorn, can be tested under full power only when all four klystrons are available.

Several additional areas of concern involve the klystron tubes. Both systems under consideration, particularly the transmission line system (Option 1), demand that the four amplifiers have matched gain and phase characteristics over the band of interest. This tube-to-tube matching could become difficult, particularly given the stretched procurement schedule for these tubes. Also, in order for the radar to operate at full output power, a VSWR of less than 1.05 to 1 must be presented to each tube. This is a difficult requirement. The feedhorn window, which must be capable of handling 1 MW CW, is also a concern. If the present material (kapton) is not suitable, an alternative must be found.

Finally, this radar transmitter would be the most complicated transmitter in the field, with the most densely populated feedcone on the 70-meter antenna, and would require special knowledge and care from the maintenance personnel at the site.

References

[1] S. A. Hovanessian, *Radar Detection and Tracking Systems*, Dedham, Massachusetts: Artech House, Chapter 11, Section 5, pp. 11–22, 1973.

[2] W. North, "Haystack Hill Long Range Imaging Radar Transmitter," Proceedings of the *13th Pulse Power Modulator Symposium*, pp. 247–253, 1978.

[3] H. R. Buchanan, *X-Band Uplink Microwave Components*, JPL Technical Report 32-1526, vol. XII, Jet Propulsion Laboratory, Pasadena, California, pp. 22–24, December 15, 1972.

[4] K. R. Goudey and A. F. Sciambi, Jr., "High Power X-Band Monopulse Tracking Feed for the Lincoln Laboratory Long Range Imaging Radar," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-26, no. 5, pp. 326–332, May 1978.

[5] D. J. Hoppe and R. M. Perez, "X-Band Resonant Ring Operation at 450 kW," *TDA Progress Report 42-93*, vol. January–March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 18–26, May 15, 1988.

Table 1. 1-MW X-band radar transmitter specifications

Parameter	Specification
Frequency	8.51 GHz
Bandwidth	20 MHz (-1 dB) 6 MHz (normal usable range)
RF output power	1 MW CW (+90 dBm)
RF stability	±0.25 dB over one planetary transmit/ receive cycle
Incidental AM	60 dB below carrier at all modulating frequencies above 1 Hz
Phase stability	10^{-15} (1000 sec)
Incidental PM (jitter)	<1° peak to peak
Transmit period	30 sec min. to 10 hr max.
Modulation:	
Phase Modulation	Biphase, 40-dB carrier suppression, dc to 20 MHz
Frequency Hopping	±2 MHz every few seconds
Frequency Ramping	±2 MHz in 200 msec
Group Delay Dispersion	10 nsec over 6-MHz bandwidth
Polarization	RHCP or LHCP (one polarization at a time; cross polarization <–25 dB)

Table 2. Characteristics of VKX-7864A X-band klystron

Parameter	Specification
Frequency	8510 MHz
Bandwidth	20 MHz (1-dB points)
Output power	250 kW min.
Beam voltage	50 kV
Beam current	11 A
Efficiency	50%
Gain (sat)	53 dB
Filament voltage	15 V
Filament current	20 A
Magnet voltage	300 V
Magnet current	20 A
Klystron weight	300 lb
Klystron height	5 ft

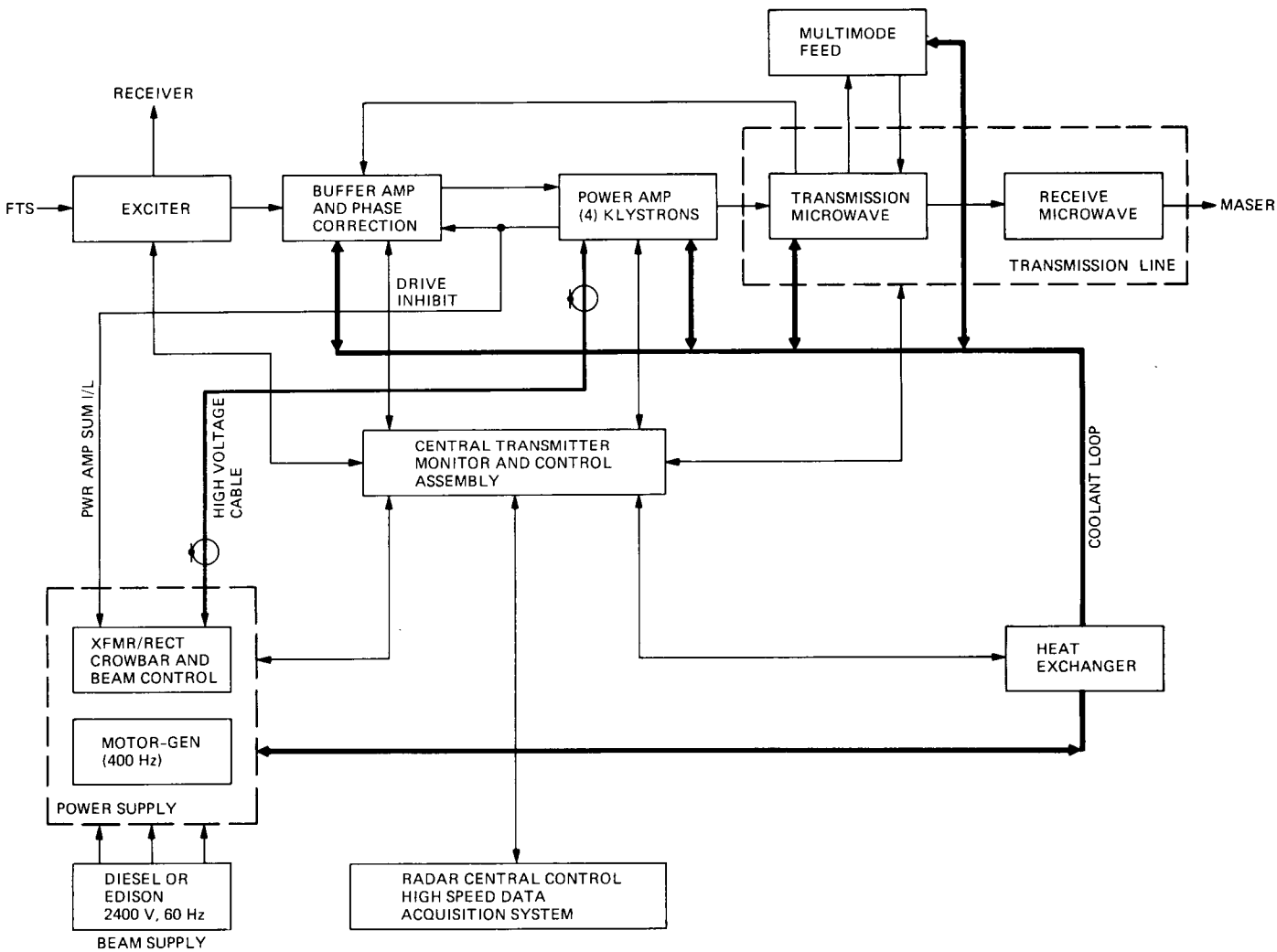
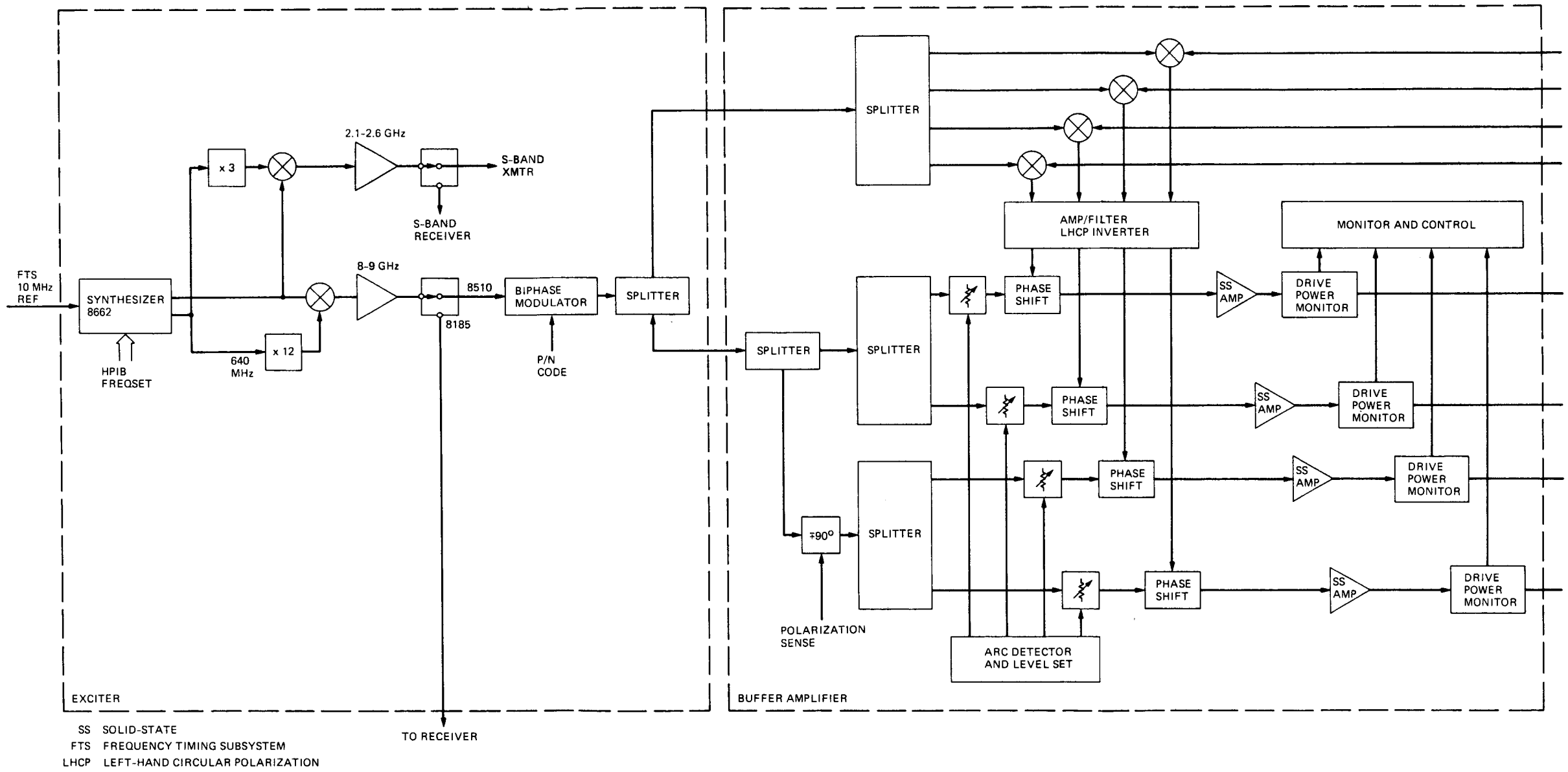


Fig. 1. 1-MW X-band radar transmitter block diagram.



SS SOLID-STATE
FTS FREQUENCY TIMING SUBSYSTEM
LHCP LEFT-HAND CIRCULAR POLARIZATION

Fig. 2. Exciter and buffer amplifier.

FOLDOUT FRAME

FOLDOUT FRAME

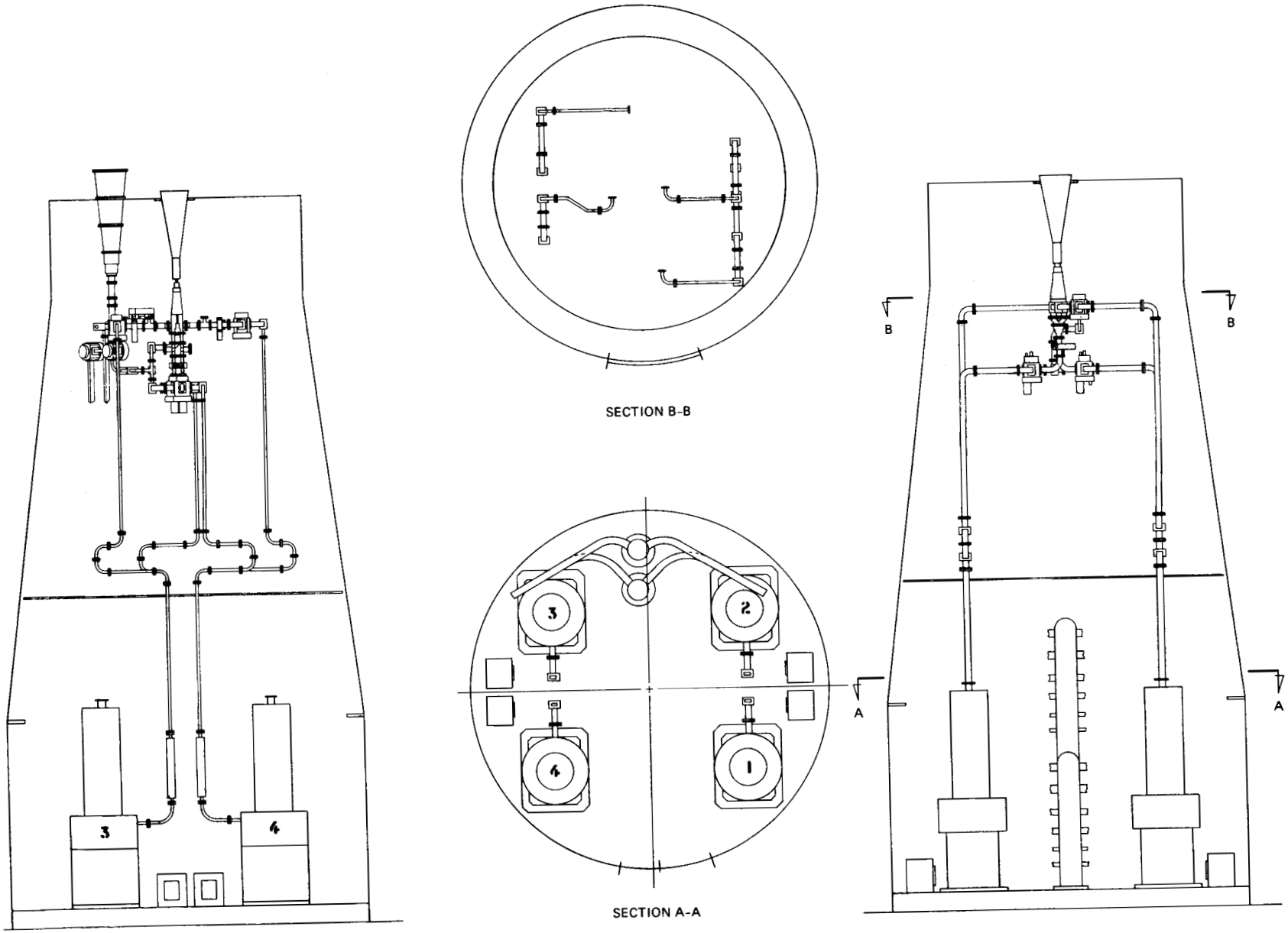


Fig. 3. Mechanical layout.

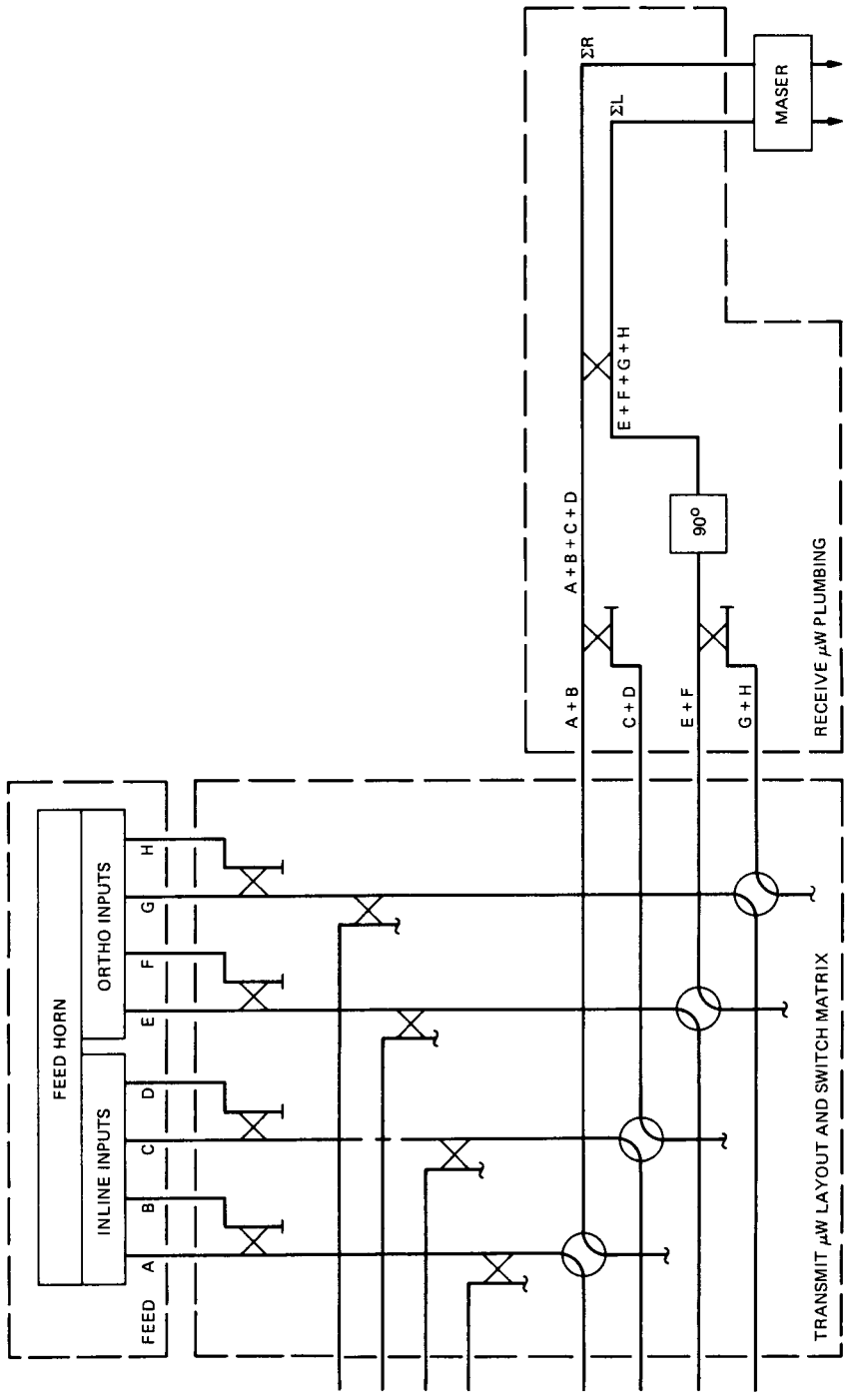


Fig. 4. Transmission line system (Option 1).

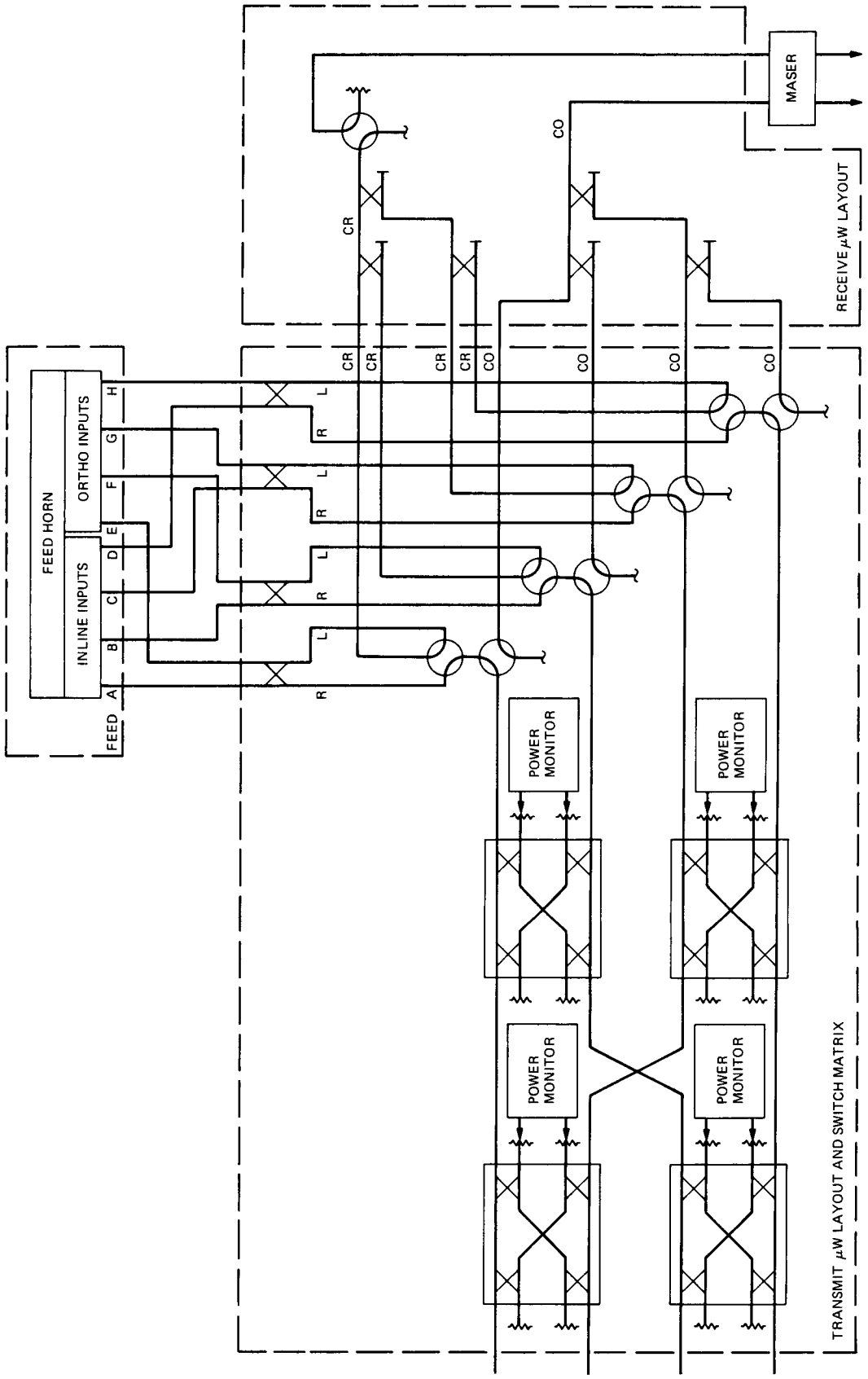


Fig. 5. Transmission line system (Option 2).

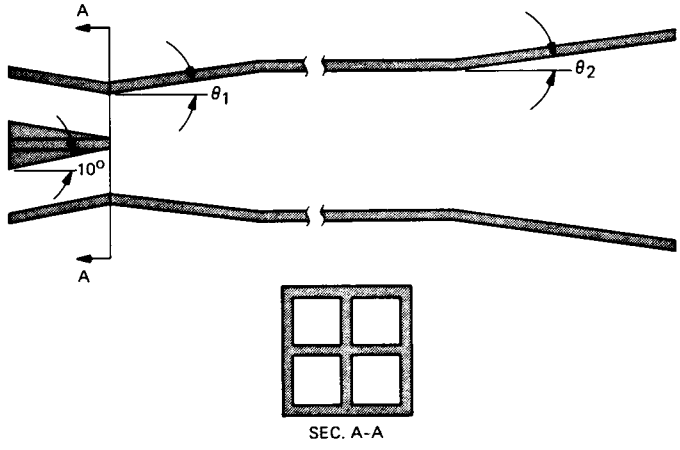


Fig. 6. 1-MW multiflare horn.

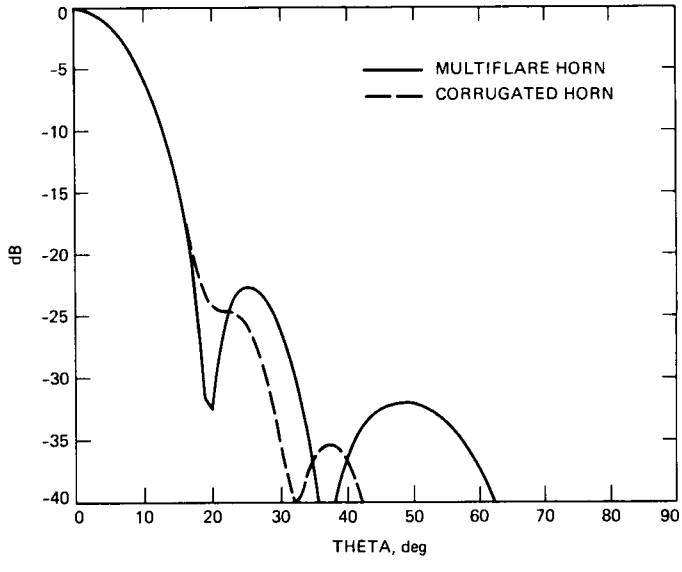


Fig. 7. 1-MW multiflare horn patterns at 8.51 GHz.

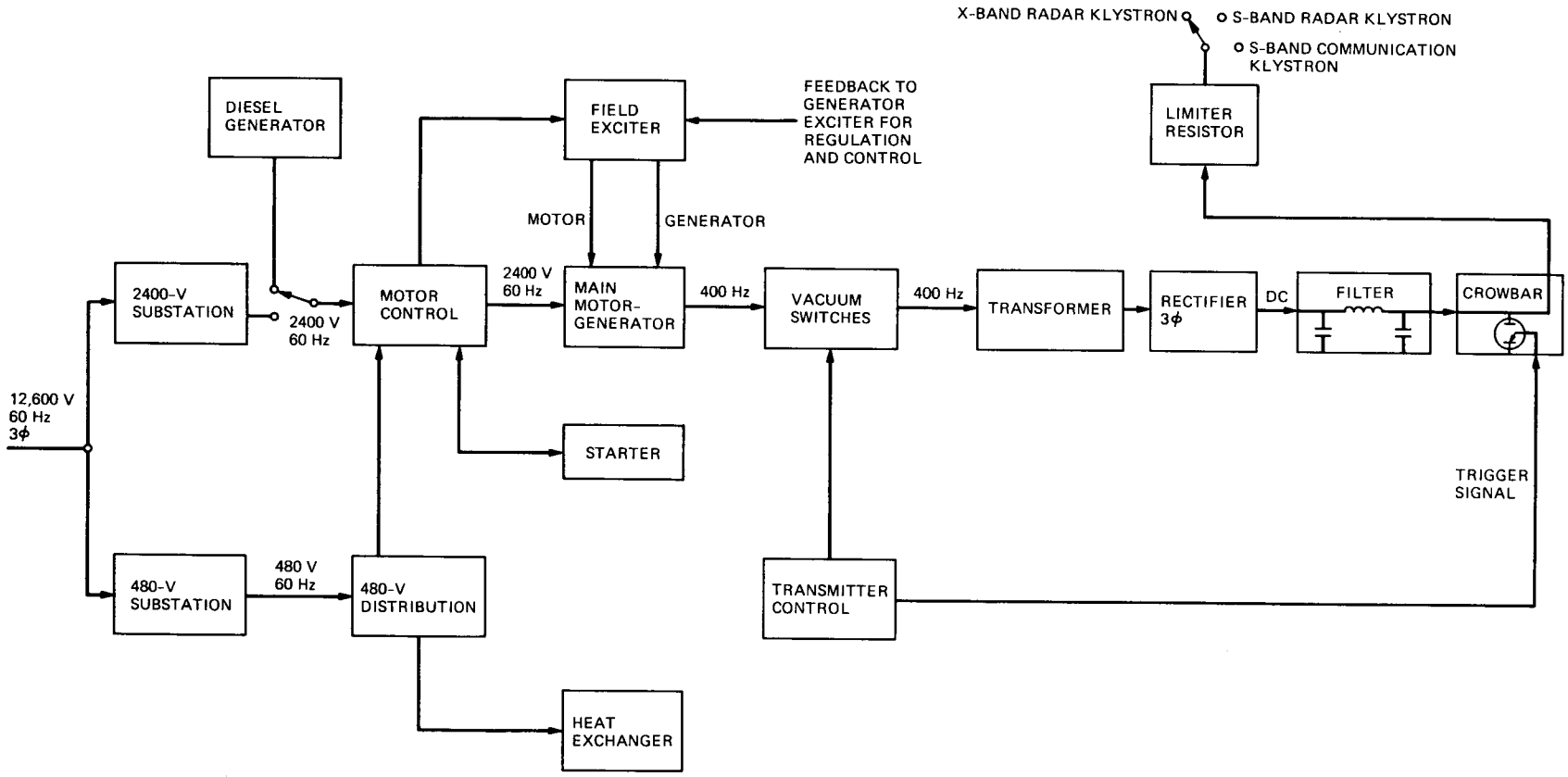


Fig. 8. Power supply block diagram.

VLA Telemetry Performance with Concatenated Coding for Voyager at Neptune

S. J. Dolinar, Jr.
Communications Systems Research Section

Current plans for supporting Voyager's encounter at Neptune include the arraying of the DSN antennas at Goldstone, California, with the National Radio Astronomy Observatory's Very Large Array (VLA) in New Mexico. Not designed as a communications antenna, the VLA's signal transmission facility suffers a disadvantage in that the received signal is subjected to a "gap" or blackout period of approximately 1.6 msec once every 5/96 sec control cycle.

Previous analyses showed that the VLA data gaps could cause disastrous performance degradation in a VLA stand-alone system and modest degradation when the VLA is arrayed equally with Goldstone. These basic conclusions were independent of whether Voyager was using its convolutional code alone or the convolutional code concatenated with its Reed-Solomon outer code.

New analysis indicates that the earlier predictions for concatenated code performance were overly pessimistic for most combinations of system parameters, including those of Voyager-VLA. The periodicity of the VLA gap cycle tends to guarantee that all Reed-Solomon codewords will receive an average share of erroneous symbols from the gaps. The number of gapped symbols is not subject to the same kind of statistical fluctuations that govern the ordinary random errors the code must also overcome. However, large deterministic fluctuations in the number of gapped symbols from codeword to codeword may occur for certain combinations of code parameters, gap cycle parameters, and data rates. In this article, several mechanisms for causing these fluctuations are identified and analyzed.

Fortunately, the Voyager-VLA parameters do not produce wild fluctuations in the number of gapped symbols from codeword to codeword. The result is graceful degradation of concatenated code performance due to the VLA gaps, even for a VLA stand-alone system. The magnitude of the deterioration at a constant concatenated code bit error rate of 10^{-5} is 0.5 dB to 0.6 dB for a VLA stand-alone system and 0.3 dB to 0.4 dB for the VLA arrayed equally with Goldstone.

Even though graceful degradation is predicted for the Voyager-VLA parameters, catastrophic degradation greater than 2 dB can occur for a VLA stand-alone system at certain non-Voyager data rates inside the range of the actual Voyager rates. Thus, it is imperative that all of the Voyager-VLA parameters be very accurately known and precisely controlled.

I. Introduction

Current plans [1] for supporting Voyager's encounter at Neptune include the arraying of the DSN antennas at Goldstone, California, with the National Radio Astronomy Observatory's Very Large Array (VLA) in New Mexico. The fully arrayed VLA operating in a stand-alone mode potentially provides about the same receiving capability as the Goldstone complex. The VLA arrayed with Goldstone would seem to offer up to two times greater data rates than Goldstone alone.

Not designed as a communications antenna, the VLA's signal transmission facility unfortunately suffers a disadvantage in that the received signal is subjected to a "gap" or blackout period of approximately 1.6 msec once every control cycle. During the blackout period, the received signal is not transmitted from the antennas to the processing facility. The control cycle is 5/96 sec (approximately 52 msec), so the blackout period constitutes about 3% of the total receiving time.

If the VLA were used in a stand-alone mode to receive uncoded data, the data received during the gaps would be irretrievably lost. The resulting bit error rate, averaged over gapped and ungapped periods, could be no better than about 1.5%, even if no errors occurred outside the gaps. Arraying the VLA with Goldstone provides some capability during the gapped periods, but the overall error rate is still at least 3% of the error rate that would prevail based on the Goldstone-only aperture without any assistance from the VLA.

The high raw error rates during the gaps can potentially be overcome by coding the data. All of Voyager's telemetry data is convolutionally encoded, and the memory and error correction capability of the convolutional code provides a mechanism for bridging small gaps in the data. Unfortunately, the convolutional code's correction capability is limited to approximately the memory length of the code, and the VLA gaps are longer than the Voyager code's memory length (6 bits) for data rates greater than 3.75 kbits/sec.

Voyager's compressed imaging data is Reed-Solomon encoded in addition to being convolutionally encoded. Each Reed-Solomon codeword consists of 255 eight-bit symbols, and blocks of four codewords are interleaved symbol by symbol. Thus, more than 8000 data bits are transmitted between the beginning and end of a codeword. At Voyager-Neptune data rates of 21.6 kbits/sec or lower, each Reed-Solomon codeword is decoded based on symbols accumulated over a minimum of seven or eight complete gap cycles, so the Reed-Solomon decoder tends to see an average mix of gapped and ungapped symbols. Since Voyager's Reed-Solomon code can correct about 6% erroneous symbols, the code can potentially withstand 3% gapped symbols with a reserve error correction capability of 3% to handle normal ungapped symbol errors.

II. Previous Analysis

The general conclusions based on the simple reasoning in Section I are largely valid, but more detailed analysis is necessary to quantify the deleterious effects of the data gaps and to detect anomalous situations when the "average" behavior is not a valid determinant of overall performance. An analysis of the effects of the VLA data gaps on Voyager's convolutionally coded and concatenated coded data was reported years ago [2], [3] when the possible use of the VLA for Voyager was first foreseen. The earlier analysis first examined the effects of the data gaps on convolutionally coded data via a full software simulation of the Viterbi decoder that accurately modeled the VLA gap cycle. Then the average Viterbi decoder error rates predicted by the simulation were used as the basis for a theoretical calculation of the performance of the convolutional/Reed-Solomon concatenated code.

The intuitive conclusions about the performance of a stand-alone VLA receiving convolutionally coded data were borne out by the simulations. As shown in Fig. 4 of [3], the decoded error rate decreases slowly as a function of signal-to-noise ratio and then flattens out at an unacceptable value around 1%. The exact limiting error rate is a function of the data rate and the duty cycle of the gap period, and it approaches 1.5% for high data rates and the VLA duty cycle of 3% gaps. When the VLA is arrayed equally with Goldstone, so that the overall signal-to-noise ratio drops by only 3 dB during the gaps, the error rate curve retains its usual character and does not approach a saturation value over the interesting range of error rates.

The earlier theoretical calculations of the Reed-Solomon code's performance in Fig. 9 of [3] show the same leveling off of error rate as a function of signal-to-noise ratio when the VLA is unassisted by Goldstone. This implies unacceptable performance for a VLA stand-alone system, no matter how high the signal-to-noise ratio at the VLA. The Reed-Solomon performance deterioration when the VLA's signal is augmented by an equal signal from Goldstone is not so dramatic, as the error rate curves again retain their usual character but drop off more slowly.

The earlier analysis reached sharply different conclusions about VLA performance with and without assistance from Goldstone. Because of the predicted potential for devastating degradation, the causes of which were not fully understood, the VLA gap analysis was reopened in order to pin down the precise error mechanisms before Voyager's Neptune encounter

next year. It was not possible to improve on the earlier analysis of the convolutional code's performance, because the simulation accurately modeled both the Viterbi decoder and the VLA gap cycle. However, a somewhat more detailed analysis of the Reed-Solomon code's performance was undertaken, and this new analysis is reported here.

The new analysis of the performance of concatenated coding yields less pessimistic conclusions about the effect of the data gaps in a VLA stand-alone system. The regularity of the gap cycle helps to eliminate the possibility of larger than average numbers of errors due to the gaps. On the average, the Reed-Solomon code's error correction capability can take care of errors during the gaps, and this average behavior is overwhelmingly likely to occur for most combinations of system parameters. On the other hand, the new analysis reveals that certain combinations of parameters are taboo if the type of ruinous degradation predicted by the old analysis is to be avoided.

III. Various Possible Analytical Approaches

There are several possible approaches to calculating the theoretical performance of the Reed-Solomon code in the presence of data gaps. Three such approaches are shown in Fig. 1. The simplest such approach, called the one-level model, was the approach used in the earlier analysis. The most complex and accurate approach, the simulated error stream model, is not feasible. The middle approach, the two-level model, is the one taken in the current analysis.

The single-level model is based on the following expression (cf. Eq. 2 of [2]) for calculating the concatenated code's bit error probability:

$$P_b = \frac{p}{\pi} \sum_{i=E+1}^{N} \frac{i}{N} \binom{N}{i} \pi^i (1-\pi)^{N-i} \qquad (1)$$

Equation (1) expresses the bit error probability P_b of the concatenated channel in terms of the bit error rate p and the Reed-Solomon symbol error rate π of the output from the Viterbi decoder. This expression assumes an $(N, N - 2E)$ Reed-Solomon code, which can correct up to E symbol errors per N-symbol codeword. In [2] and [3], the average error probabilities p and π characterizing the Viterbi decoder output were obtained by means of a detailed simulation of the Viterbi decoding process, including an accurate model of the signal-to-noise ratio fluctuations over the VLA gap cycle. The model based on Eq. (1) is termed the "single-level" model, because the Reed-Solomon error probability is calculated using single overall average values of p and π to characterize the Viterbi decoder behavior, without regard to deterministic fluctuations in p and π between the gapped and ungapped portions of a gap cycle.

The validity of Eq. (1) rests on the assumptions that successive Reed-Solomon symbol errors are independent and identically distributed. As stated in [2], independence of symbol errors is a good assumption for Voyager because Voyager's 8-bit Reed-Solomon symbols are interleaved to depth 4 and Viterbi decoder error bursts of 32 or more bits are highly unlikely for the (7, 1/2) convolutional code. On the other hand, the assumption of identically distributed symbol errors throughout a Reed-Solomon codeword should be altered to account for the deterministic periodicity of the gap cycles. Symbols occurring during the gaps have a higher error rate than symbols occurring outside the gaps.

Ideally, a set of N values of p and π should be calculated from the Viterbi decoder simulation for each possible starting "phase" of the gap cycle relative to the Reed-Solomon codeword boundaries. Equation (1) can be easily modified to allow the values of p and π to vary symbol by symbol throughout the codeword. The error rate calculated from this "multilevel" model can then be averaged over all possible relative phases of the gap cycle to obtain the overall average bit error rate.

The analysis in the present article is not based on this general multilevel model for the Viterbi decoder's output statistics. Rather, it assumes that two levels will suffice: one set of values for p and π during the ungapped portion of the gap cycle and another set of values during the gaps. Intuitively, the two-level model should become exact in the limit of very high data rates, as the widths of both the gapped and ungapped periods become long with respect to the memory length of the convolutional code. In this limit, the Viterbi decoder has a chance to settle into steady-state values of p and π both inside and outside the gaps, and the number of "transition" bits and symbols characterized by intermediate values of p and π is small relative to the number of bits and symbols characterized by the steady-state gapped and ungapped values.

Another form of the multilevel model consists of dispensing with the theoretical calculations altogether and instead feeding the simulated output of the Viterbi decoder directly into a simulation of the Reed-Solomon code's performance. Tests of this type are impractical because of the monumental amount of simulated data that must be collected before statistical confidence in the results can be obtained. For example, at an operating point of 10^{-4} Reed-Solomon codeword error probability (corresponding to a concatenated code bit error probability of about 3×10^{-6}), the average waiting time for each erroneous codeword is about 20 million

bits. Simulating enough Viterbi decoded data to produce a statistically valid sample of erroneous codewords was not feasible. However, similar end-to-end tests of the gapped VLA data have been performed using real test data from CTA-21.[1]

IV. Details of the Two-Level Model

At Voyager's data rates, the length of the ungapped portion of each gap cycle is several hundred to more than a thousand bits long, so the steady-state assumption on which the two-level model is based appears justified for the ungapped zone. The length of the gaps, however, is at most around 35 bits (or about 6 memory lengths of the convolutional code) at Voyager's highest data rate of 21.6 kbits/sec. Thus the accuracy of the two-level model is somewhat questionable for the gapped zone. However, it should still give a better prediction of concatenated code performance than the single-level model.

The two-level model used in this article is a model for the decoded output of the Viterbi decoder. The Viterbi output bit error rate is allowed to vary between two levels. The two corresponding types of errors are referred to as "gapped" errors and "ungapped" errors, respectively. Each Viterbi decoded bit is characterized by one of two bit error probabilities p_0 or p_1, and each Reed-Solomon symbol is characterized by one of two symbol error probabilities π_0 or π_1. Gapped bits and symbols have error probabilities p_0 and π_0, and ungapped bits and symbols have error probabilities p_1 and π_1.

The ungapped error probabilities p_1 and π_1 are assumed to be the steady-state Viterbi decoder output error probabilities for a decoder operating at the ungapped signal-to-noise ratio E_b/N_0, and the gapped error probabilities p_0 and π_0 are assumed to be the corresponding error probabilities for a decoder operating at the reduced signal-to-noise ratio inside the gap. The signal-to-noise ratio inside the gap is zero for the VLA stand-alone system, and for the VLA arrayed with Goldstone it is reduced from the signal-to-noise ratio outside the gap by an amount reflecting the array ratio. For equal contributions from the VLA and Goldstone, the signal-to-noise ratio reduction inside the gaps is 3 dB. Other array ratios considered in this article correspond to gap reductions of 1.5 dB and 5 dB.

A summary of the two-level model for the cases of a VLA stand-alone system and the VLA arrayed with Goldstone is shown in Table 1. Performance curves showing the relationship between the bit and symbol error probabilities and the signal-to-noise ratio for the Voyager code parameters are shown in Fig. 2[2]. The curves in this figure are taken from Fig. 3-1(a) of [4], together with the results of some new simulations of the Voyager code for signal-to-noise ratios E_b/N_0 near 0 dB or lower. The Viterbi decoder's performance in this normally uninteresting range of E_b/N_0 is relevant for the VLA gap analysis, because signal-to-noise ratios inside the gap can be 0 dB or lower whenever signal-to-noise ratios outside the gap are near the normal operating point of the decoder.

The classification of bits and symbols as "gapped" or "ungapped" is not very precise. To account for the error correction capability of the Viterbi decoder, some of the bits decoded during the gap period should be considered to be effectively the same as ungapped bits. By an argument presented in [2], the Viterbi decoder can correct exactly $K-1$ of the gapped bits when the signal-to-noise ratio is zero inside the gaps and infinite outside the gaps. Here, K is the constraint length and $K-1$ is the memory length of the convolutional code ($K=7$ for the Voyager code). In this limiting case, the effect of the convolutional code is equivalent to converting $K-1$ gapped bits (characterized by the signal-to-noise ratio inside the gap) into $K-1$ ungapped bits (characterized by the signal-to-noise ratio outside the gap). In this article, it is assumed that the convolutional code effectively accomplishes this same conversion of $K-1$ gapped bits into $K-1$ ungapped bits, even though the signal-to-noise ratios of interest are not exactly zero inside the gaps or infinite outside the gaps.

Blocks of J consecutive bits are grouped to form symbols for the Reed-Solomon code ($J=8$ for the Voyager code). A J-bit Reed-Solomon symbol is in error if any one of its bits is incorrect. Therefore, it is appropriate to classify a Reed-Solomon symbol as a gapped symbol if at least one of its J component bits is a gapped bit. Because one gapped bit can cause a whole J-bit symbol to be classified as a gapped symbol, a block of consecutive gapped bits is effectively lengthened by an average of $J-1$ bits for the purpose of calculating the number of gapped symbols. In other words, a block of B consecutive gapped bits corresponds, on the average, to $(B+J-1)/J$ gapped symbols. This effect will be analyzed more closely in Section VII.

[1] M. Varuna, "VLA Standalone Test Results for 3.6 KBPS and 7.2 KBPS Voyager Telemetry Data Rates," Interoffice Memorandum Voyager-GDSE-87-056, Jet Propulsion Laboratory, Pasadena, California, September 16, 1987.

[2] The Viterbi decoder error probabilities in Fig. 2 are plotted versus the signal-to-noise ratio E_b/N_0 for the convolutional code only. In this figure, E_b represents the signal energy per convolutionally encoded bit. In Figs. 3 through 6 and 14 through 19, which show concatenated code error probabilities, the E_b/N_0 axis represents the signal-to-noise ratio for the concatenated system, i.e., E_b is the signal energy per Reed-Solomon encoded information bit.

The actual length of the gaps is effectively reduced by about $K-1$ bits due to the error correction capability of the Viterbi decoder, but then increased by an average of $J-1$ bits by the ability of one bad bit to knock out an entire symbol. The net adjustment, $J-K$, equals just one bit for the Voyager code parameters ($J=8$, $K=7$), so any reasoning based on the physical gap length rather than the effective gap length is probably valid for the Voyager case. However, the model used in this article will keep track of these two compensating effects separately, so it can be applicable to combinations of K and J which might not cancel each other so neatly.

Since the two-level model is a model for the Viterbi decoder output statistics, it can be tested against the results of the detailed simulations conducted in [2] and [3]. The test can check whether the overall Viterbi decoder error statistics (averaged over gapped and ungapped periods) predicted from the simulations match the average of the two levels of statistics used in the two-level model. The test cannot directly check whether the two-level model is adequate for the purposes of calculating concatenated code performance, since detailed Reed-Solomon code performance simulations coupled with the Viterbi decoder simulations are not available as a benchmark. A corroboration of the two-level model at the level of the Viterbi decoded output is reported in the Appendix.

V. Code Parameters, Gap Cycle Parameters, and Data Rates

In order to apply the two-level model to the calculation of the effects of the VLA gaps on concatenated code error rates, several additional parameters need to be defined. The constraint length K or memory length $K-1$ of the convolutional code and the symbol size J of the Reed-Solomon code's symbols have already been discussed in the description of the basic model. Other code parameters that affect the concatenated code performance are the Reed-Solomon code's word length N, its error correction capability E, and its interleaving depth I. Essential parameters of the VLA gap cycle are the gap length G and the total length of the gap cycle T. The final parameter that influences the model's prediction of concatenated code performance is the data rate R. In this article, R is defined as the Viterbi decoder output bit rate. The corresponding channel symbol rate is $2R$ for Voyager's rate 1/2 convolutional code. The redundant nondata bits inserted by the Reed-Solomon code are counted toward the data rate R as defined here.

Table 2 lists the values of these essential parameters for the Voyager-VLA configuration.

VI. Conditional Concatenated Code Error Probabilities

The Reed-Solomon decoder error probability depends on the two symbol error probabilities π_0, π_1, and also on the number of symbols of both types within a Reed-Solomon codeword. Let n_0 denote the number of "gapped" symbols with error probability π_0, and n_1 the number of "ungapped" symbols with error probability π_1. The total Reed-Solomon word length is $N=n_0+n_1$. The number of gapped symbols n_0 depends on the data rate R, the gap length G and gap period T, the Reed-Solomon symbol length J and codeword length N, the convolutional code constraint length K, and the "phase" of the gap cycle relative to the Reed-Solomon codeword boundaries.

The symbol error probability for the output of the Reed-Solomon decoder can be evaluated from a generalization of Eq. (1) that accounts for the two input symbol error probability levels. The answer also depends on whether it is evaluated for a gapped symbol or an ungapped symbol. If P_{s0} and P_{s1} denote the output gapped and ungapped symbol error probabilities, respectively, then

$$P_{s0} = \sum_{\substack{0\le i\le n_0 \\ 0\le j\le n_1 \\ i+j>E}} \sum \frac{i}{n_0}\binom{n_0}{i}\pi_0^i(1-\pi_0)^{n_0-i}\binom{n_1}{j}\pi_1^j(1-\pi_1)^{n_1-j} \tag{2}$$

$$P_{s1} = \sum_{\substack{0\le i\le n_0 \\ 0\le j\le n_1 \\ i+j>E}} \sum \frac{j}{n_1}\binom{n_0}{i}\pi_0^i(1-\pi_0)^{n_0-i}\binom{n_1}{j}\pi_1^j(1-\pi_1)^{n_1-j} \tag{3}$$

The corresponding output bit error probabilities P_{b0} and P_{b1} are obtained by multiplying these expressions by the conditional probability of a bit error, given a symbol error,

$$P_{b0} = \frac{p_0}{\pi_0} P_{s0} \tag{4}$$

$$P_{b1} = \frac{p_1}{\pi_1} P_{s1} \tag{5}$$

where p_0 and p_1 are the Viterbi decoder output bit error probabilities for gapped and ungapped bits, respectively.

The overall average symbol and bit error rates P_s and P_b output from the Reed-Solomon decoder are obtained by averaging the expressions for P_{s0}, P_{s1} and P_{b0}, P_{b1} over the n_0 gapped symbols and the n_1 ungapped symbols.

$$P_s = \frac{n_0}{N} P_{s0} + \frac{n_1}{N} P_{s1} \tag{6}$$

$$P_b = \frac{n_0}{N} P_{b0} + \frac{n_1}{N} P_{b1} \tag{7}$$

The bit and symbol error probability formulas given in Eqs. (2) through (7) can be regarded as expressions for the conditional bit error probability, given knowledge of the number of gapped symbols n_0 in a codeword. The computation of this conditional error probability is a convenient intermediate step toward the eventual computation of the unconditional error probability, because it separates the performance evaluation into two reasonably distinct parts. The first part shows how sensitive the performance is to variations in the number of gapped symbols per codeword, but it can be analyzed without reference to any peculiarities of the gapping mechanism which might cause the number of gapped symbols to vary from codeword to codeword. The evaluation of the conditional bit error probability can be performed independently of many of the parameters in the problem, including the data rate R, the gap length G, the gap cycle T, and the code's interleaving depth I. The second step in the overall performance evaluation is to evaluate the interplay of these remaining parameters in determining the critical number of gapped symbols per codeword.

This separability of the overall problem, with just one crucial parameter serving to link the two parts of the analysis, is a big advantage in favor of the two-level model relative to multilevel models or a full-scale combined simulation of the Viterbi decoder and Reed-Solomon performance. Whatever exactness the simpler two-level model might lack relative to multilevel models is compensated by increased insight into what effect each of the parameters has on the overall performance.

Figures 3 through 6 show the evaluation of the Reed-Solomon conditional bit error probability for various values of the parameter n_0. Figure 3 shows the performance curves for a stand-alone VLA system, and the next three figures apply to a VLA-Goldstone array with varying contributions from each component of the array. The curves in each figure were evaluated using values of gapped and ungapped bit and symbol error probabilities, p_0, π_0, p_1, π_1, obtained from the baseline steady-state Viterbi decoder performance curves in Fig. 2. The Reed-Solomon code parameters N and E were fixed at the Voyager code's characteristics, $N = 255$ and $E = 16$.

It is instructive to examine the curves for the VLA stand-alone case. When the VLA is unassisted by Goldstone, the received data signals are totally lost during the time of the VLA gap. A total gap is characterized by zero signal-to-noise ratio and totally random steady-state Viterbi decoded bits. The gapped bit and symbol error probabilities are $p_0 = 1/2$ and $\pi_0 = (2^J - 1)/2^J$, where J is the number of bits comprising a symbol. Since $J = 8$ for the Voyager Reed-Solomon code, a good simplifying approximation is $\pi_0 \approx 1$. Substituting $\pi_0 = 1$ into Eqs. (2) and (3) leads to

$$P_{s0} = \sum_{j=E-n_0+1}^{N-n_0} \binom{N-n_0}{j} \pi_1^j (1-\pi_1)^{N-n_0-j} \tag{8}$$

$$P_{s1} = \sum_{j=E-n_0+1}^{N-n_0} \frac{j}{N-n_0} \binom{N-n_0}{j} \pi_1^j (1-\pi_1)^{N-n_0-j} \tag{9}$$

as long as n_0 is not greater than E. The error probability formula of Eq. (9) for the ungapped symbols is the same equation as for the symbol error probability under a one-level model for a code of blocklength $N - n_0$ capable of correcting $E - n_0$ errors. Thus, the effect of a total gap on the ungapped symbols is simply to use up some of the error-correcting capability of the Reed-Solomon code. The performance degradation in the ungapped zone due to the gap is equivalent to the degradation that would result from substituting a less powerful code. Figure 3 effectively shows how well a series of less and less powerful codes performs relative to the baseline performance of the Voyager code (corresponding to the curve in Fig. 3 labeled $n_0 = 0$). The performance curves deteriorate very rapidly as n_0 approaches 16, which represents a blocklength 239 code with no error correction capability. Values of n_0 greater than 16 would completely overwhelm the code.

The notion of an "equivalent" reduced-redundancy code for determining performance in the case of a total gap is useful but not completely accurate. The concatenated code's overall error statistics depend on the statistics for both ungapped symbols and gapped symbols. The error probability formula of Eq. (8) for gapped symbols equals the "equivalent" reduced-redundancy code's word error probability, which is greater than its symbol error probability. Thus, the performance degradation in the gapped zone due to the gap is worse than the degradation resulting from substituting the less powerful code. The overall average symbol error probability P_s is obtained by averaging the results of Eqs. (8) and (9) over gapped and ungapped symbols, as in Eq. (6), and hence is somewhat higher than the symbol error probability of the intuitively "equivalent" reduced-redundancy code. A similar conclusion holds for the overall average concatenated code bit error probability P_b, obtained from Eqs. (8) and (9) via Eqs. (4), (5), and (7).

VII. The Number of Gapped Symbols per Codeword

The number of bits decoded by the Viterbi decoder during each gap period is RG. The total number of Viterbi decoded bits in an entire gap cycle is RT. This corresponds to RG/r channel symbols that are received during gaps, out of RT/r channel symbols received every gap cycle if the convolutional code's rate is r ($r = 1/2$ for the Voyager code). Under the two-level model, approximately $K - 1$ (the memory length of the convolutional code) of the decoded bits during the gap period can be treated as having the same signal-to-noise ratio as ungapped bits. Thus, the effective length of each gap is reduced from RG bits to approximately $RG - K + 1$ bits due to the correction capability of the Viterbi decoder. On the other hand, the gap period is effectively lengthened by $J - 1$ bits, on the average, due to the ability of just one incorrect bit to corrupt an entire J-bit Reed-Solomon symbol. Thus, under the two-level model, the average number of gapped symbols in each N-symbol codeword is $N(RG - K + J)/RT$. This works out to an average of about eight gapped symbols per 255-symbol codeword for the Voyager-VLA parameters listed in Table 2.

The periodicity of the gap cycle tends to guarantee that every Reed-Solomon codeword receives an average number of gapped symbols. However, there are certain conditions under which this conclusion is invalid. Some codewords can get more than their share of gapped symbols, while other codewords receive fewer. The codewords receiving too many gapped symbols are drastically more prone to error, as indicated by the rapid deterioration in the performance curves in Figs. 3 to 6 as the number of gapped symbols n_0 is increased. Thus, a small number of such atypical codewords can dominate the overall error performance of the concatenated code.

A. Fluctuations Due to Symbol Edge Effects

One basic mechanism causing an uneven distribution of gapped symbols per codeword is the symbol edge effects that on the average lengthen the effective gap by $J - 1$ bits. The actual lengthening of the gap can vary from 0 bits to $2J - 2$ bits, depending on the "phase" of the gap edges relative to symbol boundaries. Figure 7 shows that each gap is lengthened by exactly $(\phi_1 + \phi_2)$ bits, where ϕ_1 and ϕ_2 are the phases of the left and right edges of the effective gap relative to Reed-Solomon symbol boundaries. Both of these phases are uniformly distributed from 0 to $J - 1$ bits for a codeword picked at random. However, the two phases are not independent of each other, and in fact they must satisfy $[(\phi_1 + \phi_2) + (RG - K + 1)] \bmod J = 0$. Despite this dependence, the average of $(\phi_1 + \phi_2)$ is always $J - 1$ bits, because the average of a sum of random variables equals the sum of the averages even when the random variables are correlated. In general, $(\phi_1 + \phi_2)$ can assume either of two values, except that when $(RG - K + 1) \bmod J = 1$, it must equal $J - 1$ bits regardless of where the symbol boundaries fall relative to the gap edges. When $(RG - K + 1) \bmod J \neq 1$, the two possible values for $(\phi_1 + \phi_2)$ are separated by exactly J bits. Thus, the actual lengthening of the gap due to symbol edge effects can be one of two values which differ by one symbol.

For example, when $(RG - K + 1) \bmod J = 0$, the two possible values for $(\phi_1 + \phi_2)$ are 0 bits and J bits, and when $(RG - K + 1) \bmod J = 2$, the two possible values are $J - 2$ bits and $2J - 2$ bits. In the first example, $(\phi_1 + \phi_2) = 0$ with probability $1/J$ and $(\phi_1 + \phi_2) = J$ with probability $1 - 1/J$. In the second example, $(\phi_1 + \phi_2) = J - 2$ with probability $1 - 1/J$ and $(\phi_1 + \phi_2) = 2J - 2$ with probability $1/J$. In both cases, the average value of $(\phi_1 + \phi_2)$ is $J - 1$ bits. However, the conditions in the second example will cause slightly poorer concatenated code performance, because the worst-case lengthening of the gap is $J - 1$ bits greater than the average lengthening. On the other hand, the performance curves for the optimal data rates, which result in $(RG - K + 1) \bmod J = 1$, will suffer no additional degradation beyond that due to the average lengthening of the gaps.

The additional degradation due to fluctuations in the lengthening of the gaps as a result of symbol edge effects is depicted in Fig. 8 as a function of the data rate R. In this figure the extra degradation is measured in terms of the worst-case number of gapped symbols per gap relative to the average number. The worst-case number of gapped symbols per gap varies periodically with the data rate between a minimum value of $(RG - K + J)/J$ gapped symbols and a maximum value of $(RG - K + J)/J + 1 - 1/J$ gapped symbols. The period of this variation is J/G. For the Voyager-VLA parameters, the worst data rates occur nominally at 5 kbits/sec, 10 kbits/sec, 15 kbits/sec, . . . , and the best data rates occur nominally at 4.375 kbits/sec, 9.375 kbits/sec, 14.375 kbits/sec, ... The exact locations of the best data rates or the worst data rates are determined under the two-level model not only by the precisely measurable value of G, but also by the assumed effective shortening of each gap by exactly $K - 1$ bits due to the error correction capability of the convolutional code. Since the effective shortening of the gap is a fuzzy quantity, the absolute locations of the best or worst data rates cannot be determined precisely. Fortunately, the variation between the best and worst data rates is relatively small, because it is equivalent to creating less than one additional gapped symbol per gap. Other mechanisms causing fluctuations in the number of gapped symbols per codeword can cause much larger effects on performance.

B. Fluctuations Due to Incomplete Gap Cycles per Codeword

A second mechanism that could cause some codewords to have an atypically large number of gapped symbols occurs when the total span of one block of interleaved codewords encompasses one more gapped section of data than another block of interleaved codewords. One block of I interleaved codewords, each consisting of N J-bit symbols, spans a continuous section of NIJ bits. If NIJ is an exact integer multiple of the gap cycle RT, than all interleaved codeword sets will include gapped symbols from exactly NIJ/RT different gap cycles. However, if NIJ/RT is not an integer, the span of an interleaved set of codewords will include a number of complete gap cycles plus a fraction of a cycle. If the extra fractional cycle includes the gap, the interleaved codeword set is "unlucky" and will suffer degraded performance, because its overall fraction of gapped symbols is larger than the nominal value of $(RG - K + J)/RT$. Other interleaved codeword sets are "lucky" and avoid the gap altogether in the fractional gap cycle, resulting in better performance than the nominal prediction. However, the overall unconditional concatenated code performance is dominated by the performance of the unlucky codeword sets, and so it is important to quantify how unlucky they can be.

Figure 9 illustrates the effects of incomplete gap cycles per interleaved codeword block. The upper picture shows the case of an unlucky codeword set with a fractional gap cycle that includes a gap, and the lower picture shows a lucky codeword set whose fractional gap cycle misses the gap. In general, an "average" interleaved codeword block includes NIJ/RT gapped sections of data, but a lucky codeword block includes only $\lfloor NIJ/RT \rfloor$, while an unlucky codeword block includes $\lfloor NIJ/RT \rfloor + 1$, where $\lfloor x \rfloor$ represents the largest integer not exceeding x. Each gapped section of data includes, on the average, $(RG - K + J)/J$ gapped symbols, which are distributed over the I interleaved codewords. The difference between the lucky codeword blocks and the unlucky ones is $(RG - K + J)/IJ$ gapped symbols per codeword per gap. The difference between an unlucky codeword block and an average one lies linearly (as a function of data rate) between 0 and $(RG - K + J)/IJ$ gapped symbols per codeword per gap, depending on how close NIJ/RT is to $\lfloor NIJ/RT \rfloor$ or to $\lfloor NIJ/RT \rfloor + 1$.

Figure 10 plots the peak-to-average concentration of gapped symbols due to incomplete gap cycles versus the data rate. The peak-to-average concentration of gapped symbols in the unlucky codewords varies between 1 and $1 + RT/NIJ$ as NIJ/RT varies between successive integer values. The concentration factor returns to 1 periodically at reciprocal data rates separated by T/NIJ, but it rises to increasing maximum values between its returns to 1. For the Voyager-VLA parameters, the maximum value of RT/NIJ is less than 1/7 even at the maximum Voyager-Neptune data rate of 21.6 kbits/sec. Thus, the overall magnitude of the degradation caused by incomplete gap cycles is limited to about one additional gapped symbol per codeword for Voyager. However, the effect of incomplete gap cycles can be very severe at higher data rates, namely data rates approaching NIJ/T (= 157 kbits/sec) or higher.

C. Fluctuations Due to Gap Cycle/Interleaving Cycle "Resonances"

A third mechanism that may cause an atypical concentration of gapped symbols in some codewords is possible "resonances" between the gap cycle and the interleaving cycle. Even if every interleaved codeword block were to experience exactly the same proportion of gapped and ungapped periods, there can be a worst-case codeword within the interleaved block which gets more than its share of gaps. In the worst conceivable case, one unlucky codeword might receive all of the interleaved block's gapped symbols, while the other $I - 1$ codewords escape with no gapped symbols at all. This would result in a worst-case peak-to-average concentration of gapped symbols in one unlucky codeword by a factor of I.

Some potential situations that may cause a concentration of gapped symbols in one unlucky codeword are illustrated in Fig. 11. In Fig. 11(a), the average effective gap period $RG - K + J$ is small enough to fit within one symbol period J, and the distance between successive gap periods (the gap cycle RT) is exactly an integer multiple of the interleaving cycle IJ. In this case, whichever of the I interleaved codewords includes the gap in its first symbol will also include all of the gaps contained within the entire span (NIJ bits) of the interleaved block. This unlucky codeword will receive I times its average share of gapped symbols, and the remaining $I - 1$ codewords will have only ungapped symbols.

Figure 11(b) illustrates a slightly different situation in which the effective gap period is still small, but the gap cycle RT is increased by enough to retard the occurrence of successive gaps by one symbol. In this case, successive gaps hit consecutive codewords, and all I codewords within the interleaved block receive a proportionate share of gapped symbols. Figure 11(c) illustrates a case in which the gap cycle RT is made slightly longer, such that it retards the occurrence of successive gaps by two symbols instead of one. Now, if the interleaving depth I is an even number, half of the interleaved codewords will get all of the gapped symbols, and the unlucky codewords will experience a peak-to-average concentration factor of 2. On the other hand, if the interleaving depth I is odd, the gaps will be distributed over all of the codewords in the interleaved block.

Figure 11(a) identifies a series of data rates R which cause major resonances between the gap cycle and the interleaving cycle. Specifically, the major resonances occur for values of R satisfying $RT \bmod IJ = 0$. Figure 11(c) shows that minor resonances can also occur when $RT \bmod IJ \neq 0$, if $(RT \bmod IJ)/J$ and I contain a common integer factor. At the major resonances, the peak-to-average concentration factor is I, while at the minor resonances the concentration factor is equal to the common integer factor of I and $(RT \bmod IJ)/J$.

The preceding conclusions about the magnitude of the peak-to-average concentration factor at the major and minor resonances are valid only for the types of cases depicted in Fig. 11, for which the gapped portion $RG - K + J$ of the total gap cycle RT is small enough to fit within one codeword symbol. Two other possible cases are illustrated in Fig. 12. In Fig. 12(a), the effective gap length $RG - K + J$ encompasses exactly two symbols, and the data rate is chosen to cause a major resonance between the gap cycle and the interleaving cycle (i.e., $RT \bmod IJ = 0$). In this case, the same two codewords are always hit by successive gaps, while the remaining $I - 2$ codewords escape the gaps altogether. The resulting peak-to-average concentration factor in this case is $I/2$.

Figure 12(b) illustrates a situation in which the effective gap length $(RG - K + J)$ is longer than one block of I interleaved symbols. The portion of the gap covering an integer multiple of IJ bits affects each of the I codewords equally, but the remaining portion of the gap covering a fraction of IJ bits afflicts one or more of the I codewords selectively. If the data rate is at a major resonance, the same codeword(s) will remain unlucky for all the gaps that occur throughout the entire span of the interleaved codeword set. The unlucky codeword(s) will receive an extra share of gapped symbols corresponding to the fractional portion of gapped bits, namely $(RG - K + J) \bmod IJ$. An average codeword should receive $(RG - K + J)/IJ$ gapped symbols from each gap, but the lucky codewords receive only $\lfloor (RG - K + J)/IJ \rfloor$, while the unlucky codewords receive $\lfloor (RG - K + J)/IJ \rfloor + 1$ (unless $(RG - K + J)$ is exactly an integer multiple of IJ). The resulting peak-to-average concentration of gapped symbols in the unlucky codewords varies between 1 and $1 + IJ/(RG - K + J)$ as $(RG - K + J)$ varies between successive integer multiples of IJ.

The location of the major resonances depends on the length of the full gap cycle RT relative to the interleaving cycle IJ, while the magnitude of the performance deterioration at each major resonance depends on the effective length $(RG - K + J)$ of the gapped portion of a gap cycle relative to the interleaving cycle IJ. These two effects are depicted separately in Fig. 13. The variation of the peak-to-average concentration of gapped symbols is shown as a smooth curve, while the resonance locations are shown as sharp lines. The peak-to-average concentrations depicted by the smooth curve are valid only at or near the resonance locations. The resonances themselves are of nonzero width, but they are very narrow if the number of symbols N per codeword is large.

Peak-to-average concentration factors of 1.5 and greater are common at major resonances within the range of Voyager's data rates. The maximum peak-to-average concentration factor varies between 4 and 2 for data rates between 4.375 and 19.375 kbits/sec. A peak-to-average concentration factor of 1.5 corresponds to 12 gapped symbols per worst-case codeword rather than the average of 8. Figure 3 shows the potential for catastrophic performance degradation of a VLA standalone system when the number of gapped symbols per codeword reaches 12 or 16 or higher. Thus, it is important that the precise Voyager data rates miss the location of the narrow resonances. Fortunately, this is the case. Table 3 lists the data rates causing major resonances in the range from 3.75 to 22.5 kbits/sec. Even though one of the Voyager data rates (21.6 kbits/sec) appears to fall perilously close to a major resonance, the unconditional performance evaluations in the next section show that the small separation is sufficient to avoid resonant degradation.

VIII. Unconditional Concatenated Code Error Probability Curves for the Voyager-VLA Parameters

The number of gapped symbols per codeword, n_0, was evaluated as a function of the relative phase between the gap cycle and the codeword, assuming the Voyager-VLA parameter values listed in Table 2. This evaluation simultaneously takes into account all of the types of fluctuations identified in the previous section. It was found that the Voyager data rates are all nonresonant, in the sense that the worst-case value of n_0 was 9 or 10 and not 12, 16, or higher.

The overall unconditional concatenated code bit error rate is obtained by averaging the bit error rate in Eq. (7),

$$\bar{P}_b = \mathbf{E}\{P_b\} \tag{10}$$

where $\mathbf{E}\{\cdot\}$ represents an average over the possible values of n_0. Unconditional concatenated code bit error rates calculated from Eq. (10) are plotted in Figs. 14 through 17 for the same four VLA-Goldstone array ratios considered in Figs. 3 through 6.

The concatenated code performance curves in Figs. 14 through 17 are virtually identical for the three Voyager data rates shown, R = 7.2 kbits/sec, 14.4 kbits/sec, and 21.6 kbits/sec. Performance is very slightly improved at the higher rates. This negligible difference is attributable to the net effective

lengthening of the gap portion of the gap cycle by $J - K = 1$ bit. This net lengthening constitutes a larger fraction of the total gap cycle at lower data rates than at higher rates, and hence the gap is predicted to affect the lower data rates slightly more adversely. However, as pointed out earlier, the net effective lengthening of the gap is the result of two compensating effects which almost cancel each other. The convolutional code's error correction capability makes the gaps look shorter at the low data rates, while edge effects due to entire Reed-Solomon symbols getting wiped out by one erroneous bit make the gaps look shorter at the high data rates. The model for both of these effects is fuzzy enough that the predicted tiny performance improvement with increasing data rate is not significant. A more appropriate conclusion is that the predicted concatenated code performance is virtually independent of the Voyager data rate over the range $R = 7.2$ to 21.6 kbits/sec.

The performance curves in Figs. 14 through 17 are essentially identical to the $n_0 = 9$ conditional error rate curves in Figs. 3 through 6. This indicates that the unconditional error rate is almost completely determined by the error rate for codewords with the worst-case number of gapped symbols. At a constant performance level of 10^{-5} bit error rate, the net effect of the VLA gaps is to require 0.5 dB to 0.6 dB more signal-to-noise ratio for the VLA stand-alone system relative to an ungapped system. The net cost of the gaps when the VLA is arrayed equally with Goldstone (3-dB gaps) is 0.3 dB to 0.4 dB. When the VLA's contribution to the array is about twice Goldstone's (5-dB gaps), the net cost of the gaps is almost the same as for the VLA stand-alone system, 0.5 dB. When Goldstone's contribution is about twice the VLA's (1.5-dB gaps), the net cost of the gaps shrinks to 0.1 dB to 0.2 dB.

The modest amount of deterioration in the concatenated code performance due to the VLA gaps is a consequence of the reserve error correction capability of the Reed-Solomon code relative to the average number of gapped symbols, coupled with the fortuitous choice of nonresonant data rates for Voyager. A remarkable example of catastrophic performance degradation due to a resonant data rate is shown in Fig. 18. The resonant data rate, $R = 21.504$ kbits/sec, differs from one of the important Voyager rates by less than 0.5%, and yet this case suffers an additional performance degradation of around 2 dB for the VLA stand-alone system. The explanation is that 21.504 kbits/sec is a resonant data rate with a worst-case number of gapped symbols per codeword equal to 16 (see Table 3), which exhausts the error correction capacity of the Reed-Solomon code.

The effects of a resonant data rate are not so pronounced when the VLA is arrayed with Goldstone. Figure 19 shows concatenated code performance for the same resonant data rate considered in Fig. 18, but for the case of an equal VLA-Goldstone array ratio (3-dB gaps). The resonant data rate performance is only 0.1 dB to 0.2 dB worse than the performance for the nonresonant Voyager data rate at a required bit error probability of 10^{-5}. In fact, the concatenated code's performance at high bit error rates ($>10^{-3}$) is slightly better at the resonant rate than at the nonresonant rate. The reason for the dramatically improved performance is that, even though the Viterbi decoder's error rate during the 3-dB gaps is truly bad, it does not decode completely random bits as it does when the gaps are totally devoid of received data. If the Viterbi decoder manages to decode a few gapped symbols correctly, every correctly decoded gapped symbol adds one symbol's worth of reserve correction capacity to a Reed-Solomon decoder that would otherwise be operating with essentially no reserve capacity at all.

The main lesson to be drawn from Figs. 14 through 19 is that the Voyager data rates and the VLA gap cycle parameters must be very accurately known and precisely controlled in order to avoid a disastrous resonance that would ruin the performance of a VLA stand-alone system. If this can be done, the overall concatenated code performance degradation due to the VLA gaps can be limited to about 0.5 to 0.6 dB. When the VLA and Goldstone are arrayed in equal ratio, the nominal degradation is reduced to 0.3 to 0.4 dB, but, just as significantly, the extra degradation at a resonant data rate is only a few more tenths of a dB. Thus, the necessity to avoid a resonant data rate is not quite so critical if the VLA is arrayed with Goldstone.

IX. Summary

Voyager's Reed-Solomon outer code has sufficient error correction capacity to withstand the average number of erroneous symbols caused by the VLA data gaps. Of course, the code's reserve capacity for correcting ordinary random errors not caused by the gaps is diminished and the overall concatenated code performance is slightly degraded relative to that of an ungapped receiving system.

The periodicity of the VLA gap cycle tends to distribute an average number of gapped symbols to every Reed-Solomon codeword. However, several mechanisms were identified which can cause the actual number of gapped symbols to deviate from its benign average value for some unlucky codewords. These fluctuations are important because the overall performance of the concatenated code is dominated by its performance for the unluckiest codewords, which receive the worst-case number of gapped symbols. The mechanism causing the most serious fluctuations within the range of the Voyager-

VLA parameters is resonances between the VLA gap cycle and the Reed-Solomon codeword interleaving cycle. At many resonances within the range of Voyager's data rates, the number of gapped symbols included in unlucky codewords is 1.5 to 4 times higher than the average number. The performance degradation at these resonant data rates can be catastrophic, especially for a VLA stand-alone system, as seen in Fig. 18. Fortunately, the resonances are very narrow and none of the actual Voyager data rates falls disastrously near a resonant rate. However, the existence of these catastrophic resonant rates inside the range of the actual Voyager data rates underscores the importance of accurately knowing and precisely controlling all of the relevant code parameters, gap cycle parameters, and data rates for the Voyager-VLA system.

References

[1] J. W. Layland and D. W. Brown, "Planning for VLA/DSN Arrayed Support to the Voyager at Neptune," *TDA Progress Report 42-82*, vol. April–June 1985, pp. 125–135, Jet Propulsion Laboratory, Pasadena, California, August 15, 1985.

[2] L. J. Deutsch, "The Performance of VLA as a Telemetry Receiver for Voyager Planetary Encounters," *TDA Progress Report 42-71*, vol. July–September 1982, pp. 27–39, Jet Propulsion Laboratory, Pasadena, California, November 14, 1982.

[3] L. J. Deutsch, "An Update on the Use of the VLA for Telemetry Reception," *TDA Progress Report 42-72*, vol. October–December 1982, pp. 51–60, Jet Propulsion Laboratory, Pasadena, California, February 15, 1983.

[4] R. L. Miller, L. J. Deutsch, and S. A. Butman, *On the Error Statistics of Viterbi Decoding and the Performance of Concatenated Codes*, JPL Publication 81-9, Jet Propulsion Laboratory, Pasadena, California, September 1, 1981.

Table 1. The two-level model for the Viterbi decoder output statistics

	VLA-Goldstone array		VLA stand-alone system	
	Inside the gaps	Outside the gaps	Inside the gaps	Outside the gaps
Bit error rate	p_0	p_1	$p_0 \approx 1/2$	p_1
Symbol error rate	π_0	π_1	$\pi_0 \approx 1$	π_1
Signal-to-noise ratio	$\rho\, E_b/N_0$*	E_b/N_0	0	E_b/N_0

*Values of the array ratio considered in this article are: $10 \log_{10} \rho = 1.5$ dB, 3 dB, 5 dB (as well as $\rho = 0$ for the VLA stand-alone case).

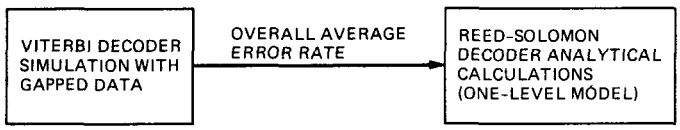
Table 2. Code parameters, gap cycle parameters, and data rates

		General Case	Voyager-VLA Case
Convolutional code parameters			
Constraint length		K	7
Memory length		$K-1$	6
Code rate		r	1/2
Reed-Solomon code parameters			
Symbol size		J	8
Codeword size		N	255
Error correction capability		E	16
Code rate		$1-2E/N$	223/255
Interleaving depth		I	4
Gap cycle parameters			
Gap length		G	1.6 msec
Total gap cycle length		T	5/96 sec
Data rate		R	21.6 kbits/sec
(Viterbi decoder bit rate)			14.4 kbits/sec
			7.2 kbits/sec
			3.6 kbits/sec

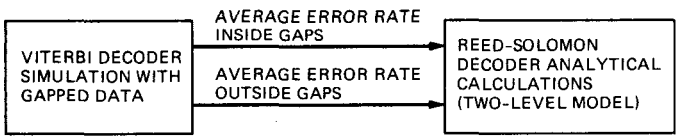
Table 3. Data rates causing major resonances between the gap cycle and the interleaving cycle

Resonant data rate (kbits/sec)	Worst-case number of gapped symbols per codeword
4.3008	37
4.9152	32
5.5296	29
6.1440	26
6.7584	24
7.3728	22
7.9872	20
8.6016	19
9.2160	17
9.8304	16
10.4448	15
11.0592	15
11.6736	14
12.2880	13
12.9024	13
13.5168	12
14.1312	12
14.7456	11
15.3600	11
15.9744	10
16.5888	10
17.2032	10
17.8176	9
18.4320	9
19.0464	18
19.6608	16
20.2752	16
20.8896	16
21.5040	16
22.1184	16

(a) ONE-LEVEL AVERAGE ERROR RATE MODEL ([2], [3])



(b) TWO-LEVEL AVERAGE ERROR RATE MODEL (PRESENT ARTICLE)



(c) SIMULATED ERROR STREAM MODEL (NOT PRACTICAL)

Fig. 1. Various analytical approaches to modeling concatenated coding with gapped data.

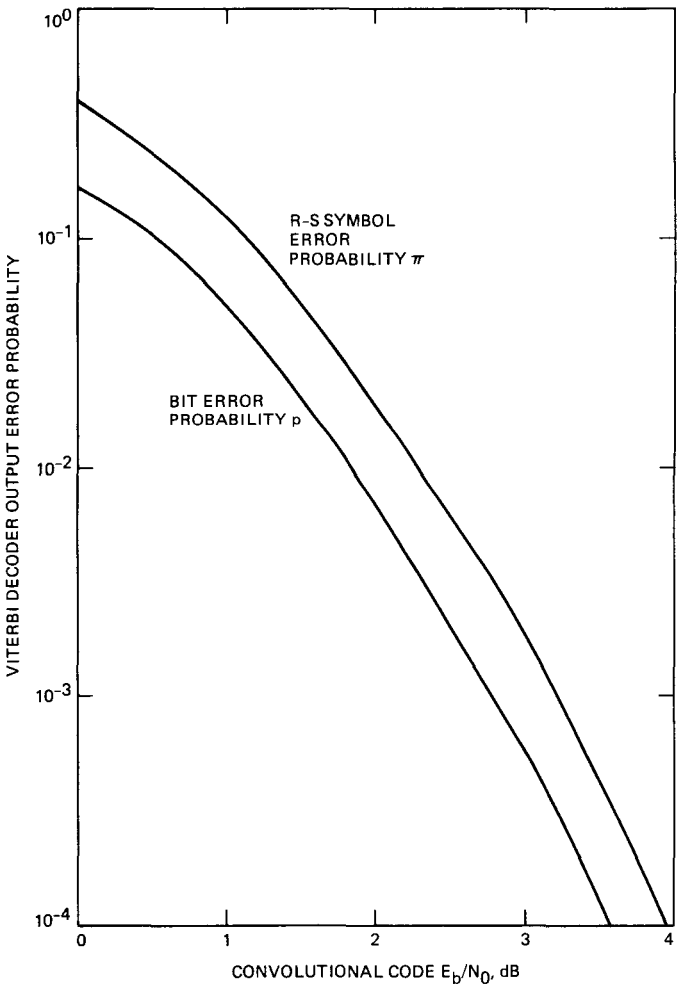


Fig. 2. Viterbi decoder output error probabilities for ungapped system (taken from [5], Fig. 3-1).

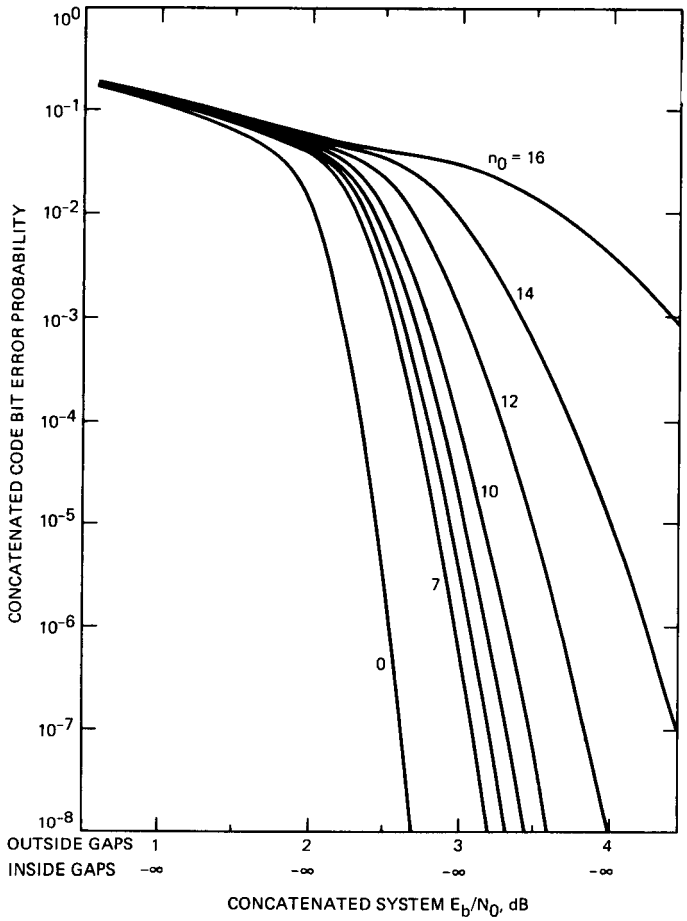


Fig. 3. Conditional concatenated performance for VLA stand-alone system as a function of n_0 = number of gapped symbols per codeword.

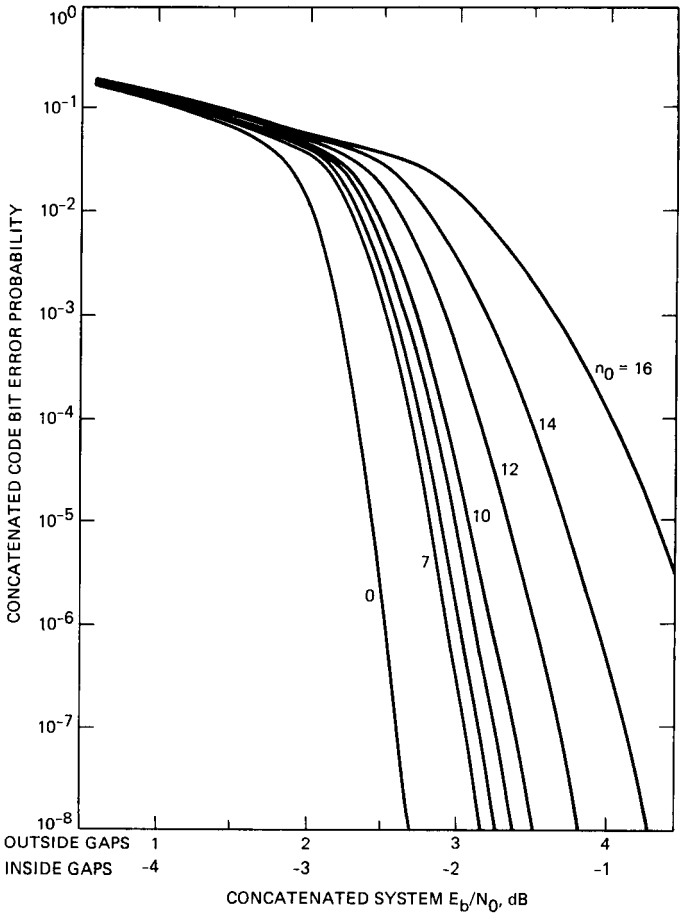


Fig. 4. Conditional concatenated code performance for VLA arrayed unequally with Goldstone (5-dB gaps) as a function of n_0 = number of gapped symbols per codeword.

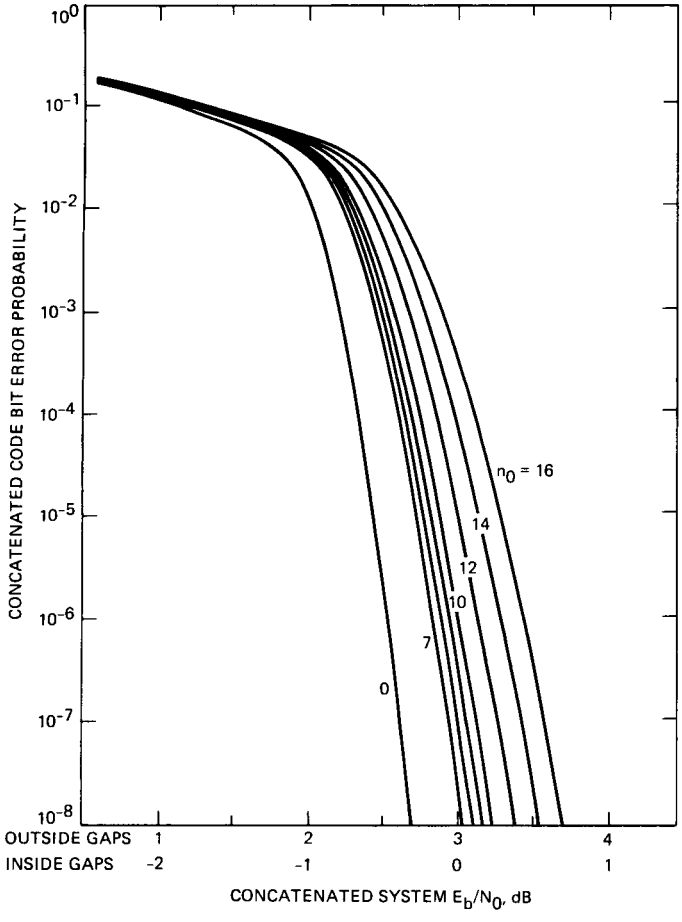


Fig. 5. Conditional concatenated code performance for VLA arrayed equally with Goldstone (3-dB gaps) as a function of n_0 = number of gapped symbols per codeword.

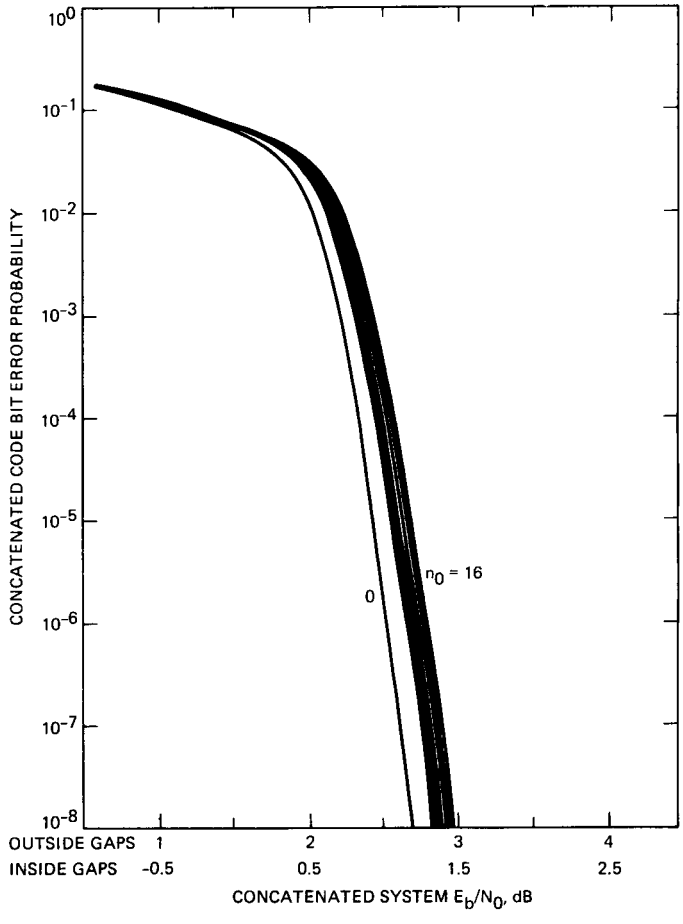


Fig. 6. Conditional concatenated code performance for VLA arrayed unequally with Goldstone (1.5-dB gaps) as a function of n_0 = number of gapped symbols per codeword.

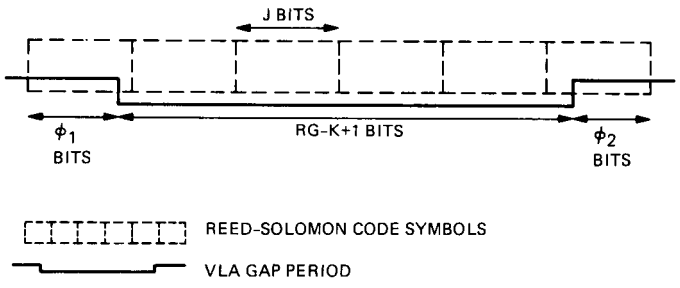


Fig. 7. Symbol edge effects.

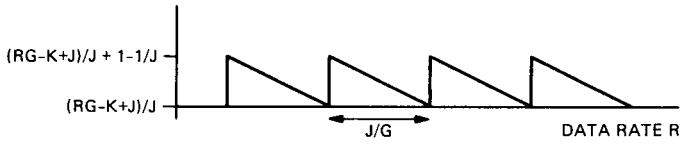


Fig. 8. Fluctuations in the number of gapped symbols per gap due to worst-case symbol edge effects.

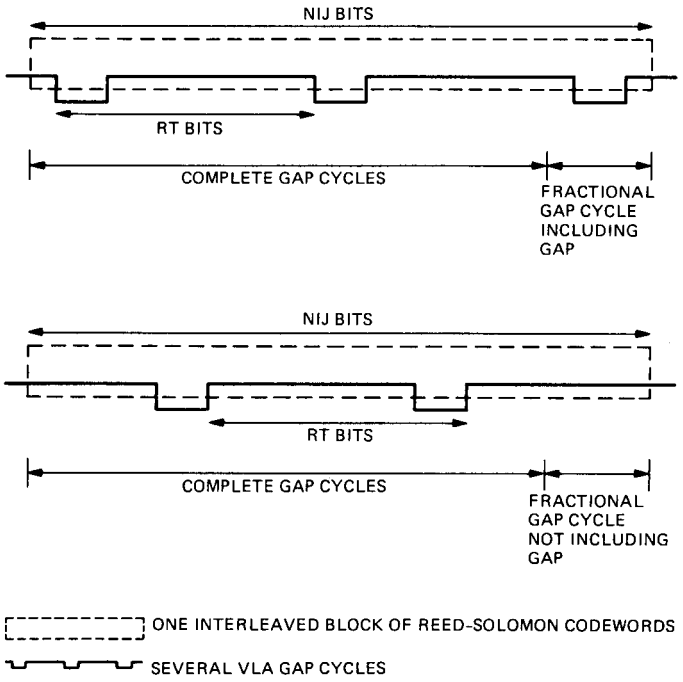


Fig. 9. Incomplete gap cycle effects.

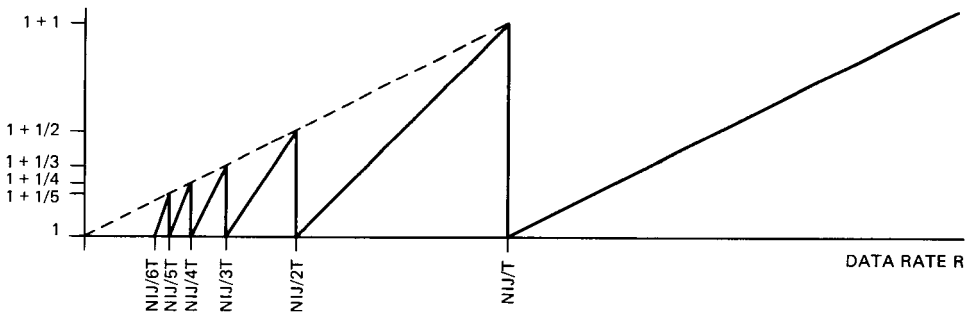


Fig. 10. Fluctuations in the peak-to-average concentration of gapped symbols per codeword due to incomplete gap cycles per interleaved block of codewords.

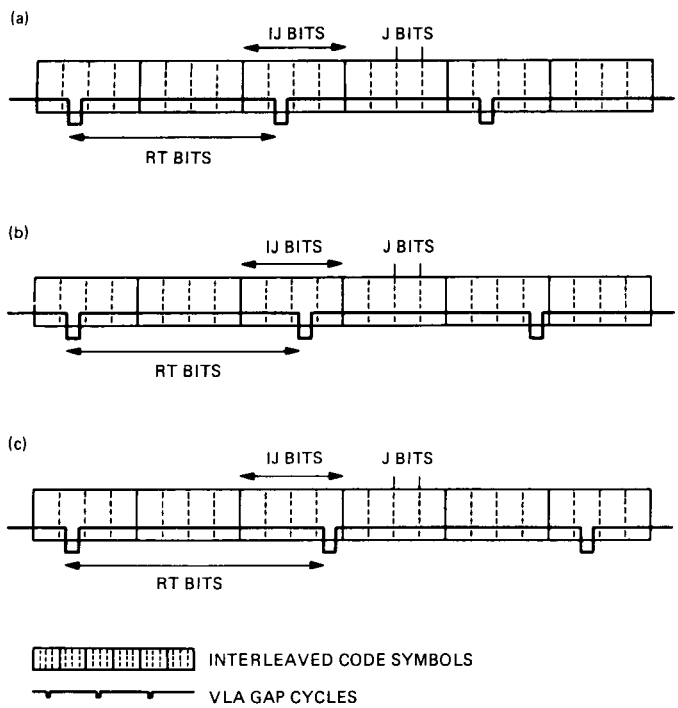


Fig. 11. Gap-cycle/interleaving-cycle resonance effects for average effective gap lengths up to one symbol.

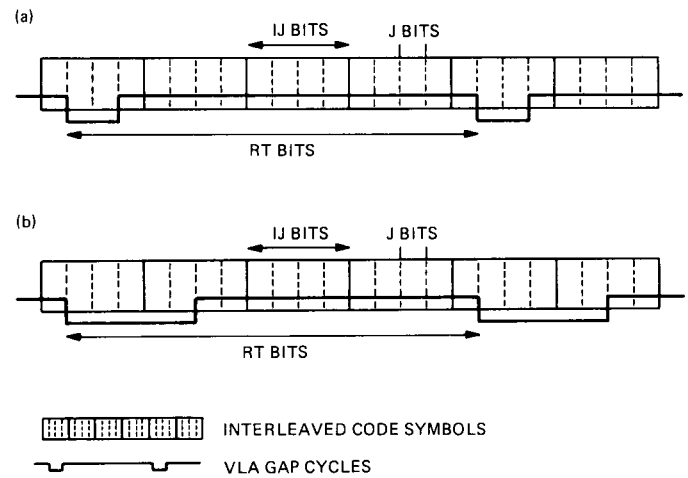


Fig. 12. Gap-cycle/interleaving-cycle resonance effects for average effective gap lengths greater than one symbol.

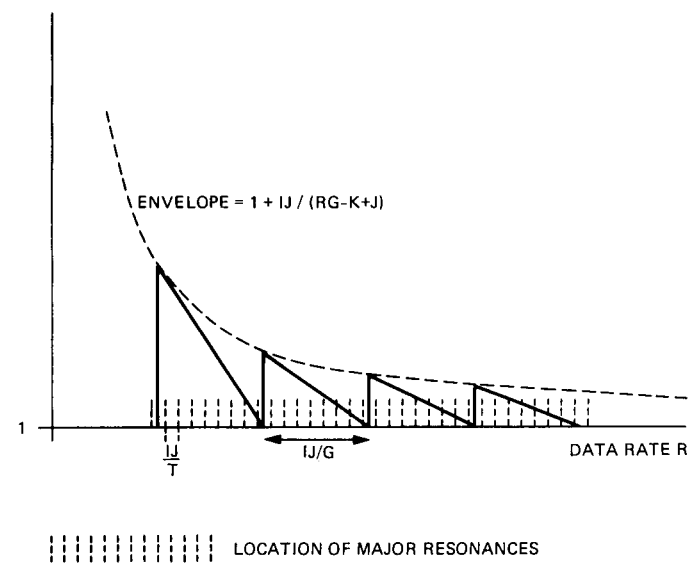


Fig. 13. Fluctuations in the peak-to-average concentration of gapped symbols per codeword due to gap-cycle/interleaving-cycle resonances. For clarity, figure depicts resonance spacing corresponding to T/G = 8. Actual T/G = 32.55 for the VLA parameters.

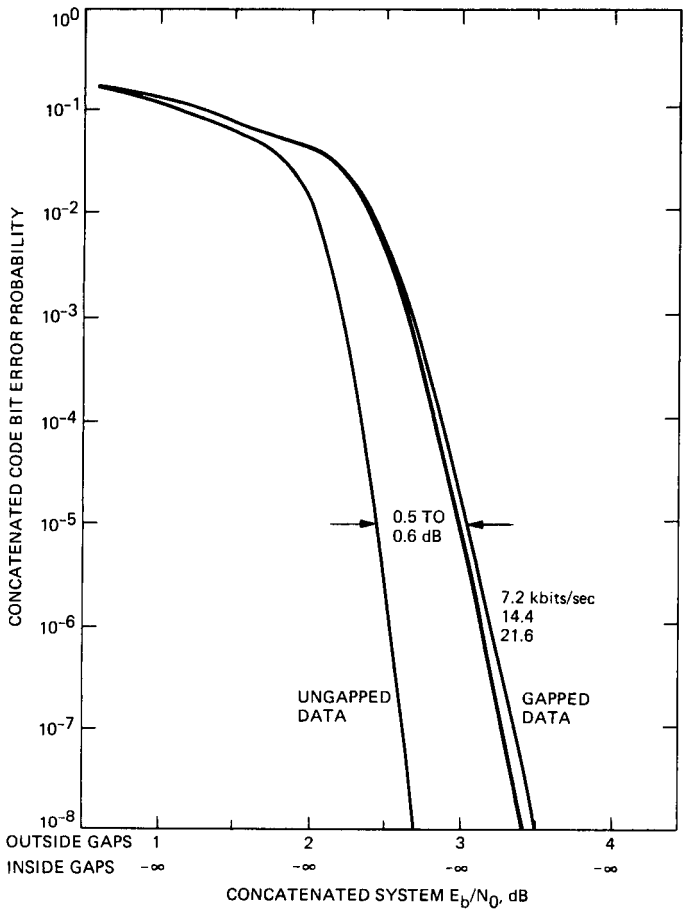


Fig. 14. Unconditional concatenated code performance at Voyager-Neptune data rates for VLA stand-alone system.

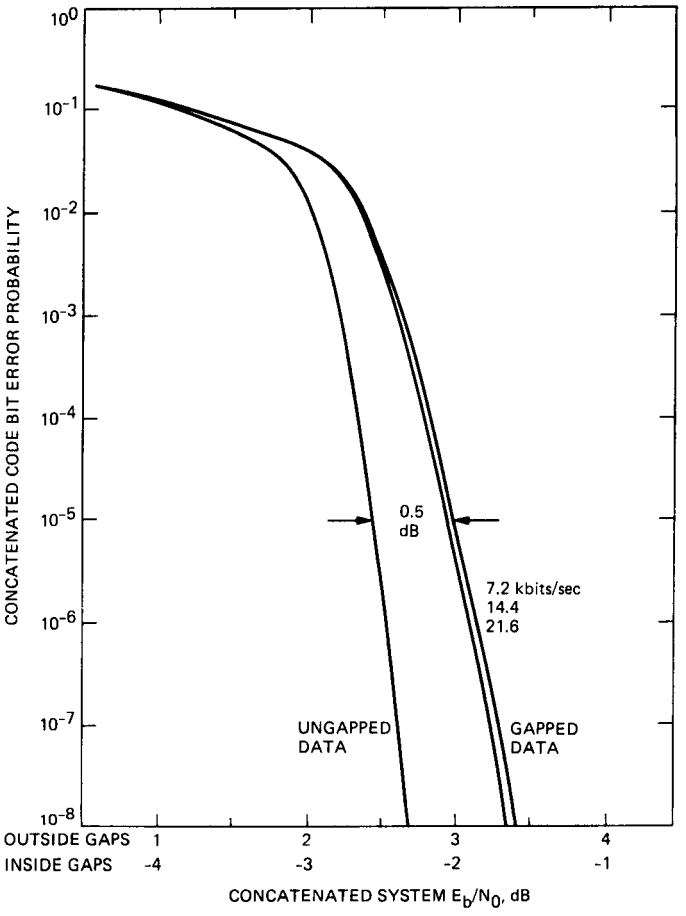


Fig. 15. Unconditional concatenated code performance at Voyager-Neptune data rates for VLA arrayed unequally with Goldstone (5-dB gaps).

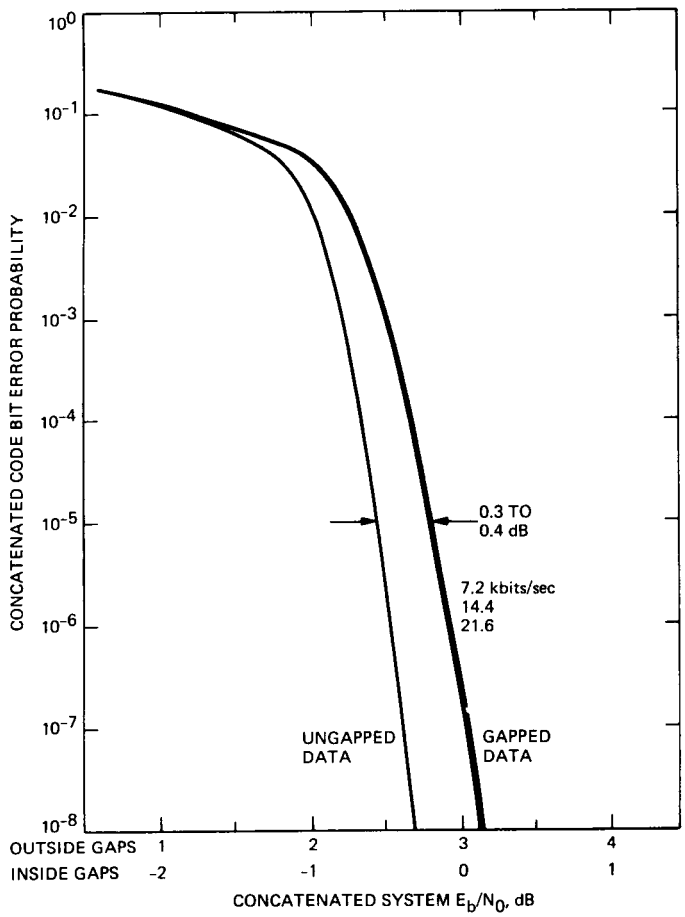


Fig. 16. Unconditional concatenated code performance at Voyager-Neptune data rates for VLA arrayed equally with Goldstone (3-dB gaps).

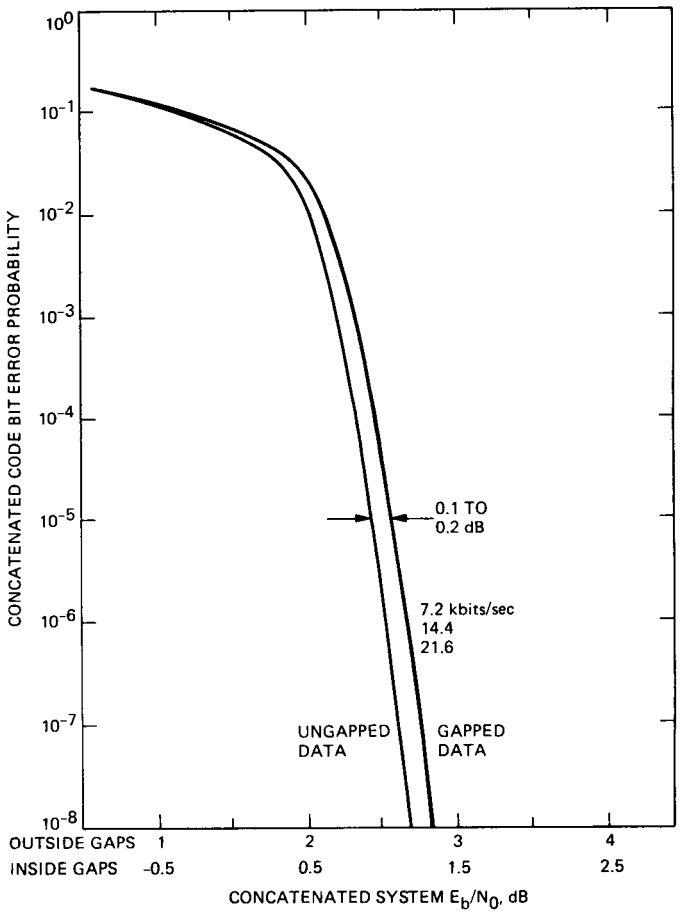


Fig. 17. Unconditional concatenated code performance at Voyager-Neptune data rates for VLA arrayed unequally with Goldstone (1.5-dB gaps).

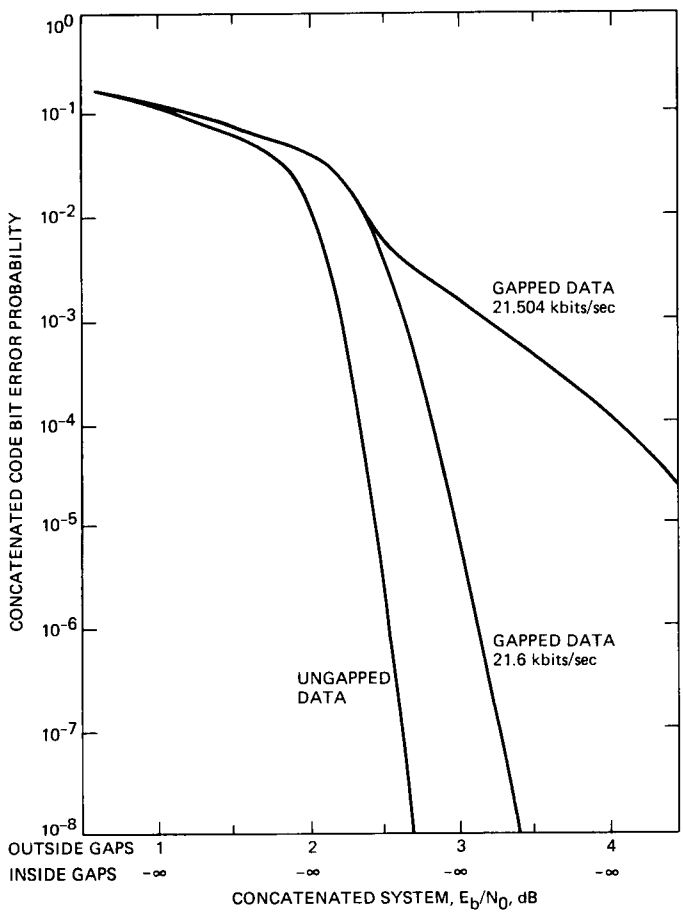


Fig. 18. Comparison of unconditional concatenated code performance at a Voyager-Neptune data rate and a nearby resonant data rate for VLA stand-alone system.

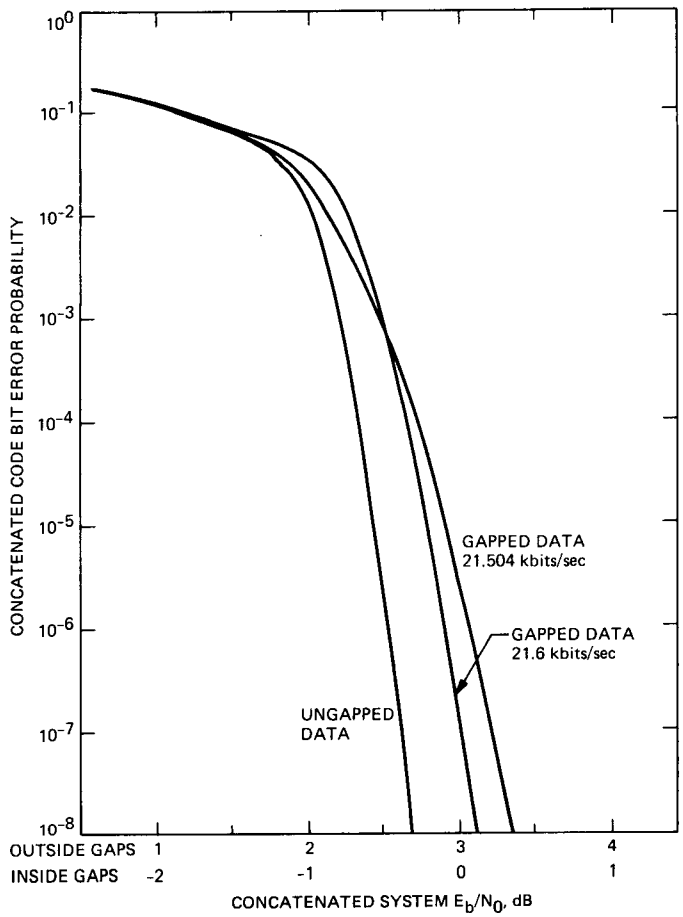


Fig. 19. Comparison of unconditional concatenated code performance at a Voyager-Neptune data rate and a nearby resonant data rate for VLA arrayed equally with Goldstone (3-dB gaps).

Appendix
Corroboration of the Two-Level Model

The two-level model for the Viterbi decoder output statistics is an ad hoc model that was chosen for simplicity. It is relatively easy to analyze, and it affords easy separability of the analysis into a conditional evaluation of code performance and a computation of the gap cycle's effect on the symbols in any given codeword. At the same time, it allows the gapped and ungapped portions of the gap cycle to be treated differently, not just characterized by overall cycle averages as in the even simpler one-level model.

The validity of the two-level model for the purposes of calculating concatenated code performance could be completely confirmed only by end-to-end tests or simulations of the entire concatenated code. Such simulations are impractical and were not performed. However, a partial corroboration of the model's accuracy can be obtained by comparing the average of the gapped and ungapped Viterbi decoder error rates predicted by the two-level models to the overall average Viterbi decoder error rates obtained from simulations of the Viterbi decoder operating over many VLA gap cycles. If the two-level model is accurate, the following equation should hold:

$$p = \frac{RG - K + 1}{RT} p_0 + \left(1 - \frac{RG - K + 1}{RT}\right) p_1 \qquad \text{(A-1)}$$

Here, p_0 and p_1 are the gapped and ungapped bit error probabilities defined in Table 1 and Fig. 2 and used in Eqs. (4) and (5), and p is the simulated bit error rate (averaged over many gap cycles) used in Eq. (1) of this article and plotted in Figs. 3 and 4 of [2].

Figures A-1 and A-2 compare the left sides and right sides of the purported Eq. (A-1) for the VLA stand-alone case and for the VLA arrayed equally with Goldstone.[1] Fairly good agreement is obtained. Even the substantial variation of performance with data rate is somewhat accurately predicted by the two-level model. The agreement is best for the VLA stand-alone case. The gap conditions in this case are more nearly equal to the assumption of zero signal-to-noise ratio inside the gaps and infinite signal-to-noise ratio outside the gaps, from which the argument for effectively shortening the gaps by $K - 1$ bits was derived.

A second form of justification for the two-level model is obtained by examining its validity at the end points. It was already stated in Section IV that the model is exactly correct in the extreme limit of zero signal-to-noise ratio inside the gaps and infinite signal-to-noise ratio outside the gaps. These extreme limits represent the maximum possible difference between the characteristics of the gapped and ungapped portions of the gap cycle. The opposite extreme occurs as the difference between the gapped zone and the ungapped zone goes to zero and the two signal-to-noise ratios become equal. The two-level model is obviously exactly correct in this extreme limit also, because Eqs. (2) through (7) reduce to Eq. (1) when $p_0 = p_1 = p$ and $\pi_0 = \pi_1 = \pi$. These end-point arguments do not directly confirm the validity of the two-level model in the intermediate regions, but nonetheless they inspire confidence that it should not be too far wrong.

[1] For comparison with [2], VLA gaps are assumed to be 1 msec instead of 1.6 msec for the curves in Figs. A-1 and A-2.

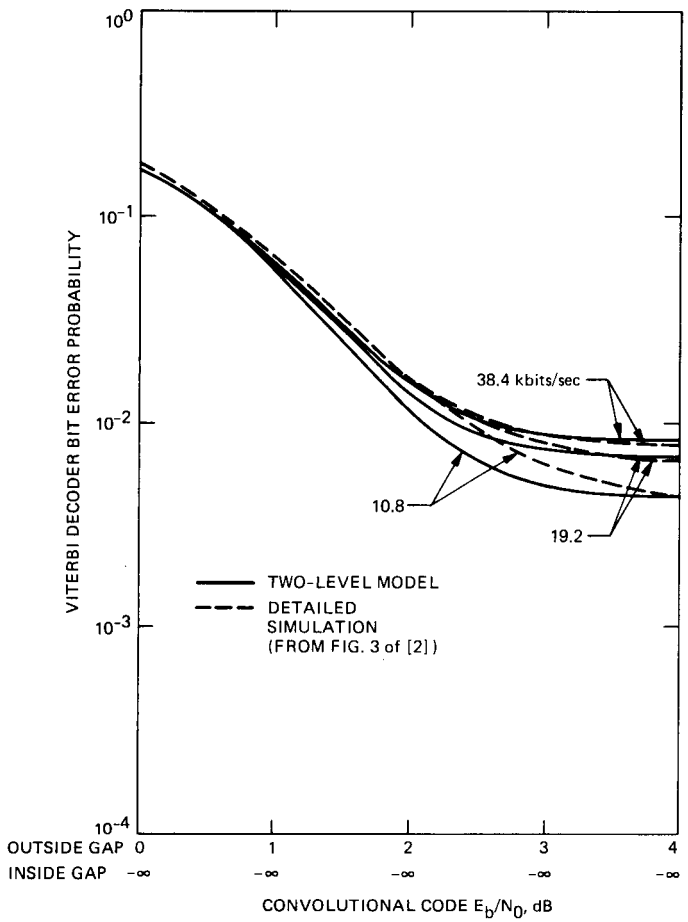


Fig. A-1. Comparison of Viterbi decoder performance predicted by two-level model and by detailed simulation for VLA stand-alone system. (Note: For comparison with [2], VLA gaps are assumed to be 1 msec instead of 1.6 msec for these curves.)

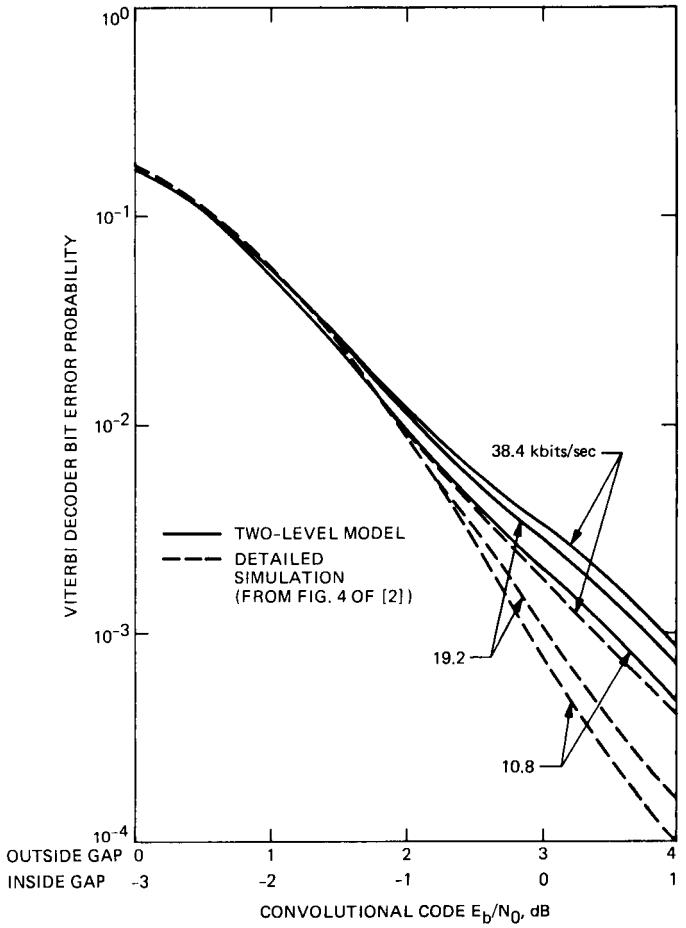


Fig. A-2. Comparison of Viterbi decoder performance predicted by two-level model and by detailed simulation for VLA arrayed equally with Goldstone (3-dB gaps). (Note: For comparison with [2], VLA gaps are assumed to be 1 msec instead of 1.6 msec for these curves.)

A Long Constraint Length VLSI Viterbi Decoder for the DSN

J. Statman, G. Zimmerman, F. Pollara, and O. Collins
Communications Systems Research Section

A new Viterbi decoder, capable of decoding convolutional codes with constraint lengths up to 15, is under development for the DSN. The objective is to complete a prototype of this decoder by late 1990, and demonstrate its performance using the (15, 1/4) encoder in Galileo. The decoder is expected to provide 1 dB to 2 dB improvement in bit SNR, compared to the present (7,1/2) code and existing Maximum-Likelihood Convolutional Decoder (MCD). The new decoder will be fully programmable for any code up to constraint length 15, and code rate 1/2 to 1/6. This article describes the decoder architecture and top-level design.

I. Introduction

The DSN uses concatentated codes to reduce the Bit Error Rate (BER) on the telemetry channel from deep space probes to the DSN complexes. Standard coding, as used for the Voyager mission, consists of an outer (255,223) Reed Solomon (RS) code and an inner convolutional code with constraint length $K = 7$, and code rate 1/2. Decoding is accomplished by a Maximum-Likelihood Convolutional Decoder (MCD), followed by an RS decoder. A typical telemetry chain is shown in Fig. 1. Performance of this coding scheme is well understood [1],[2].

Recently [3], new convolutional codes have been discovered that provide a "2-dB coding gain" over existing codes. The highest gain, 2.11 dB, is achieved by using a (15,1/6) convolutional code, concatenated with a (1023,959) RS code. Using (15,1/6) convolutional codes with a (255,223) RS code results in an estimated coding gain of 1.8 dB. This gain can be realized by building a new Viterbi decoder for the inner code, and using the existing RS decoder. Hence, employing the newly discovered convolutional codes can result in relatively inexpensive improvement in DSN telemetry performance.

To demonstrate the new codes, an encoder for a (15,1/4) convolutional code is being added to the Galileo spacecraft. A rate 1/4 code is used instead of a rate 1/6 code, because of the limited bandwidth available on the Galileo modulator. This encoder, shown in Fig. 2, requires only a small number of parts (20 integrated circuits and 60 discrete components) and thus has a minimal impact on spacecraft complexity. A prototype decoder is being developed, capable of decoding Galileo data, but also of accepting other codes such as DSN standard codes and (15,1/6) convolutional codes. Figure 3 shows the BER versus bit SNR, for various coding schemes, with a predicted coding gain for the Galileo experiment of 1.5 dB.

The complexity of a Viterbi decoder depends on three key parameters: constraint length (i.e., degree of the generating polynomials), code rate (i.e., reciprocal of the number of encoded symbols transmitted for each information bit), and information data rate. The major complexity driver is constraint length, since the amount of hardware is roughly proportional to the number of states, which is $2^{(K-1)}$, where K is the constraint length. Hence a decoder for $K = 15$ is approximately 256 times more complex than a decoder for $K = 7$. Such a

complex decoder can be built with current VLSI technology within reasonable size limitations.

This paper describes the prototype decoder. Section 2 outlines system requirements, and Section 3 describes the top-level design. Section 4 describes in detail the architecture of the processor assembly, the unit performing the actual decoding.

II. Decoder Requirements

Requirements for the decoder can be separated into three categories:

(1) Performance. The decoder will process convolutionally coded data with constraint length up to 15 (programmable) and code rate 1/2 to 1/6 (programmable). Data rate must meet Galileo requirement (134.4 Kbit/sec), with a goal of 1 Mbit/sec. The decoder will utilize a synchronization pattern, if it is present in the uncoded data stream, to support node synch. In addition, an external node synch input will be available.

(2) Interfaces. The decoder will provide DSN interfaces, for testing in CTA 21 and for integration into DSN complexes. At a minimum these include symbol input from the Symbol Synchronizer Assembly (SSA) or the Base-Band Assembly (BBA), decoded information bits to the Frame Synchronization Subsystem (FSS), and interfaces to station monitor and control.

(3) Testability. The decoder will include testing capability for both stand-alone tests and DSN compatibility tests. In the stand-alone test, the decoder will generate a pre-programmed information bit sequence, encode it according to the desired convolutional code, add a programmable amount of white Gaussian noise to the symbols, pass the noisy symbols to the decoder proper (processor assembly), compare the decoded bits to the original sequence, and compute BER, in real time. The decoder will also provide GO or NO GO indication to the operator. For DSN compatibility testing the decoder will receive a symbol stream, and an un-encoded bit stream, decode the symbols, and compute BER.

Additional requirements concerning operating environment, size, power consumption, reliability, fault testing, and maintainability exist, but are not discussed here.[1]

[1] J. Statman, "Draft Task Plan for Large Constraint Length VLSI Viterbi Decoder," JPL IOM 331-87.5-241 (internal document), December 28, 1987.

III. Top-level Design

A functional block diagram of the decoder is shown in Fig. 4. The following is an overview of these blocks:

(1) Processor Assembly. This is the "heart" of the decoder. It consists of 256 identical VLSI chips that perform the maximum-likelihood decoding of the incoming symbol sequence. In addition, this assembly includes path memory, metric normalization circuitry, and the applicable computer, timing, and control interfaces.

(2) Simulator Assembly. The simulator assembly generates a noisy symbol sequence in three steps. First, an information sequence is generated. Next, this sequence is encoded using the appropriate convolutional encoder. Finally, a measured amount of noise is added. In addition, the simulator assembly sends the uncoded information sequence to the comparator assembly, to enable performance evaluation.

(3) Comparator Assembly. This assembly receives "true" information bits from either the simulator assembly or from an external input, and decoded bits from the processor assembly. It aligns the sequences and collects BER data.

(4) Node Synch Assembly. The node synch assembly derives node synch either from the rate of metric increase, from an embedded synch pattern, or from an external source.

(5) Erasure Signal Generator. This is an option under consideration. It is based on an algorithm [4] that compares the incoming symbols to an encoded version of the decoded information bits, to determine probable burst-error locations.

(6) SSA Interface. This module converts the signal coming from the SSA to signals compatible with the decoder. The two key operations are adjustment of voltage levels and removal of additional sign inversions added by some encoders.

(7) FSS Interface. This module sends the decoded bits to the FSS, similar to the existing MCD output.

(8) Other DSN Interfaces. More DSN interfaces are under definition. Options are interface to the Telemetry Processor Assembly (TPA) and interfaces to future DSN data network via Small Computer Standard Interface (SCSI) bus for data transfer, and General Purpose Instrumentation bus (GPIB) for monitor and control.

(9) Computer-Controller-Timing. The computer-controller-timing coordinates the modules described above by providing command, control, and monitor operations during initialization and decoding. In addition, it generates all the required timing signals and allows for extensive stand-alone testing.

The prototype decoder packaging approach is to provide for easy transfer from prototype packaging to a DSN-ready system. Standard DSN packaging techniques are used where possible. The baseline package is in two drawers, mountable in a 19-inch rack. The first drawer is based on a MULTIBUS I card cage and includes all the assemblies and external interfaces, except for the processor assembly. The second drawer includes the processor assembly.

IV. Processor Assembly Architecture

The architecture presented here is for a particular implementation of the Viterbi decoder. We start by reviewing several basic definitions and algorithms that are used elsewhere in this article. It is not intended as a Viterbi decoder tutorial, and the interested reader may read references [1], [2], [5], [6] for further information.

The Viterbi decoder tries to find the best possible match between a stream of received symbols and a path through a state trellis. The processing is sequential, i.e., using the set of symbols corresponding to a single information bit, the decoder progresses from one time-slice through the trellis to the next, while updating its decision on the most likely path and the resulting decoded bits. For a code with constraint length K, the number of states is 2^{K-1}, so in the $K = 15$ decoder there are 16384 states.

We assume here that the code rate is $1/n$. This implies that each state is connected to two preceding states and to two succeeding states, depending on whether the preceding and succeeding information bits are 0 or 1. In fact, it is convenient to organize the states in butterflies (so called because the graph of associated arithmetic resembles a butterfly). Each butterfly contains two states, and has inputs from two other butterflies and outputs to two other butterflies. For $K = 15$, the 16384 states are organized in 8192 butterflies.

The data exchanged between the butterflies are accumulated metrics. These metrics represent the probability of trellis paths, i.e., the lower the accumulated metric, the more likely is the path. There is one accumulated metric per state, or two per butterfly. Accumulated metrics are computed inside the butterfly. For each set of symbols corresponding to an information bit, the butterflies add the existing accumulated metric to the metric associated with the new symbols (so called "branch metric"), resulting in new accumulated metrics. As time passes, accumulated metrics grow, so periodically they are reduced down, or normalized.

A. Basic Trade-Offs

Several implementation choices were made and are documented below. First, the 8192 butterflies can be implemented using serial or parallel architectures, or with a hybrid serial-parallel approach. In a serial architecture, a single physical butterfly processor performs all 8192 butterflies, sequentially. In a parallel architecture, 8192 physical butterflies are used. In a hybrid approach, n physical butterflies are used, each sequencing through $8192/n$ butterflies. The *fully parallel architecture* was chosen.

Next, a choice of arithmetic method is made. The arithmetic operations include addition, subtraction, and comparisons between metrics. The decoder uses integer arithmetic and performs *bit-serial arithmetic*, or bit-by-bit operations. In this approach, the metrics (represented by 8- to 18-bit numbers) are sent serially, on a single wire, LSB to MSB. A separate TDA Progress Report article is under preparation, describing the bit-serial versus parallel arithmetic trade-offs.

Next, the method for decoder graph partitioning is selected. Butterfly interconnection can be represented by a graph with 8192 nodes, where each node corresponds to a butterfly. Each node has inputs from two other nodes and outputs to two other nodes. The partitioning selected will be described in detail in a future progress report. It is a two-level partitioning of the graph, where the first-level subgraphs correspond to printed circuit boards, while secondlevel subgraphs correspond to VLSI chips. Key features of the partitioning are (a) the graph is split among 16 *identical* boards, each with 16 *identical* VLSI chips, leading to easy implementation, (b) any Viterbi decoder of constraint length K can be built by wiring together $2^{(K-7)}$ of these chips or $2^{(K-11)}$ of these boards, and (c) the number of wires between boards and chips is relatively small.

B. Processor Assembly Elements

The processor assembly, shown in Fig. 5, consists of six major functions:

(1) Symbol Conversion. The symbols arriving into the processor assembly are 8-bit 2's complement quantities, arriving at the rate of one symbol per symbol clock. The symbol conversion module buffers the symbols into blocks that correspond to information bits (using the node synch signal), converts the symbols into sign-magnitude values, and rearranges the symbols for bit-serial transmission to the butterflies. It also

computes the sum of the magnitudes of the six symbols and transmits it to butterflies, bit-serially, LSB first.

(2) Butterflies. The butterflies are the core of the decoder. As shown in Fig. 6, each butterfly consists of two main blocks: an Add-Compare-Select (ACS) unit and a Metric Computer. The ACS uses four adders to add branch metrics to accumulated metrics, then compares the sums to select two of them for further transmission. The metric computer uses a set of adders to compute a weighted sum of the received symbols. Both the ACS and the metric computer are mathematically specified below. The complete decoder for $K = 15$ has 8192 butterflies, 32 butterflies per VLSI chip.

(3) Metric Exchange. The metric exchange function is performed by the interconnections between butterflies. Some of the metrics are exchanged inside the VLSI chip, while others are sent via wires between chips and in a backplane. All transmitted metrics must be kept aligned, i.e., the ith bit of transmitted metric is present on all metric exchange wires at the same clock period, regardless of the form of this connection.

(4) Traceback Memory. After each butterfly completes the ACS operation it sends two bits to the traceback memory. These two bits per butterfly (computed once per information bit) represent the results of the two ACS select operations. The traceback memory can be viewed as a matrix where one dimension is the number of states, 16384, and the other dimension corresponds to time, and has $3*7*K$ entries. For $K = 15$, the memory has at least 16384 *3*7*15 bits, or approximately 640 Kbytes.

(5) Traceback Processor. The traceback processor reads and writes the traceback memory to produce decoded bits [4].[2]

(6) Normalization Processor. The normalization processor monitors several accumulated metrics. When any of these metrics exceeds a computer-selected threshold, a normalization command is issued to the butterflies, to be executed during the next information bit time.

C. Butterfly Mathematical Representation and Implementation

The following paragraphs describe the equations of the ACS and the metric computer unit.

[2]F. Pollara and H. Shao, "Memory Management in Traceback Viterbi Decoders," JPL IOM 331-87.2-242 (internal document), February 12, 1987.

1. Add-Compare-Select. A diagram of an ACS is shown in Fig. 7. The accumulated metrics (16-bits) from neighboring states $i0$ and $i1$, which were previously computed in some other ACS unit, are added bit-serially to the branch metrics b_{00}, b_{01}, b_{10}, and b_{11}, provided by the metric computer unit. The operation produces the sums:

$$s_{00} = m_{i0} + b_{00}$$

$$s_{10} = m_{i1} + b_{10}$$

$$s_{01} = m_{i0} + b_{01}$$

$$s_{11} = m_{i1} + b_{11}$$

These sums are shifted into four shift registers and the smaller sum of each pair is selected by the comparators, as follows:

$$\text{if } (s_{00} < s_{10}),\ m_{j0} = s_{00} \text{ and } bit0 = 0,$$

$$\text{otherwise } m_{j0} = s_{10} \text{ and } bit0 = 1$$

$$\text{if } (s_{01} < s_{11}),\ m_{j1} = s_{01} \text{ and } bit1 = 0,$$

$$\text{otherwise } m_{j1} = s_{11} \text{ and } bit1 = 1$$

Here, m_{j0} and m_{j1} are the output accumulated metrics, and $bit0$ and $bit1$ are the results of the decisions, sent to the traceback memory.

2. Branch Metric Computer. The branch metrics are computed in the metric computer (Fig. 8) from the six received symbols, $r_0 \ldots r_5$. The implementation here is slightly different from that found in the literature, resulting in reduction of the dynamic range of the branch metrics by a factor of 2 [4]. Let r_i be represented as sign and magnitude binary numbers, as follows:

$$r_i = (s_i, r_{i6}, r_{i5}, r_{i4}, r_{i3}, r_{i2}, r_{i1}, r_{i0})$$

where s_i are the sign bits. Let $(e_0, e_1, \ldots, e_5)$ be the label assigned to the butterfly at initialization. This label is the output of an encoder making one of the state transitions of the butterfly. Because the generator polynomials of the codes considered have leading and trailing ones, each butterfly has only two possible branch metrics, and they sum to a constant (for a fixed set of symbols). Let

$$c_i = e_i \oplus s_i \qquad i = 0, \ldots, 5$$

then

$$b_{00} = \sum_{i \in A} \hat{r}_i$$

where A is the set of all i's such that $c_i = 1$, and $\hat{r}_i$ are the magnitudes of the r_i's. Also, $b_{10} = x - b_{00}$, where x is

$$x = \sum_{i=0}^{5} \hat{r}_i$$

Finally, for codes with a leading and trailing one in the generator polynomials, $b_{11} = b_{00}$ and $b_{01} = b_{10}$. This condition is met for our codes with $K = 15$. If $K < 15$, we have $b_{11} = b_{10}$ and $b_{01} = b_{00}$.

V. Conclusions

A new DSN Viterbi decoder is under development that benefits from two recent Advanced Systems developments: the successful search for long constraint length codes which yield a "2-dB coding gain," and the VLSI expertise in the Communications Systems Research Section. The top-level design, mathematical characterization, and functional specifications have been completed. The decoder is expected to be ready for testing using a Galileo encoder by late 1990.

References

[1] J. H. Yuen, *Deep Space Communications Systems Engineering*, New York: Plenum Press, 1983.

[2] R. L. Miller, L. J. Deutsch, and S. A. Butman, *On the Error Statistics of Viterbi Decoding and the Performance of Concatenated Codes*, JPL Publication 81-58, Jet Propulsion Laboratory, Pasadena, California, September 1981.

[3] J. H. Yuen and Q. D. Vo, "In Search of a 2-dB Coding Gain," *TDA Progress Report 42-83*, vol. July–September 1985, Jet Propulsion Laboratory, Pasadena, California, pp. 26–33, November 15, 1985.

[4] O. Collins, Ph.D. Thesis, California Institute of Technology, in preparation.

[5] G. C. Clark and J. B. Cain, *Error-Correcting Coding for Digital Communications*, New York: Plenum Press, 1981.

[6] R. J. McEliece, *The Theory of Information and Coding*, Massachusetts: Cambridge Press, 1977.

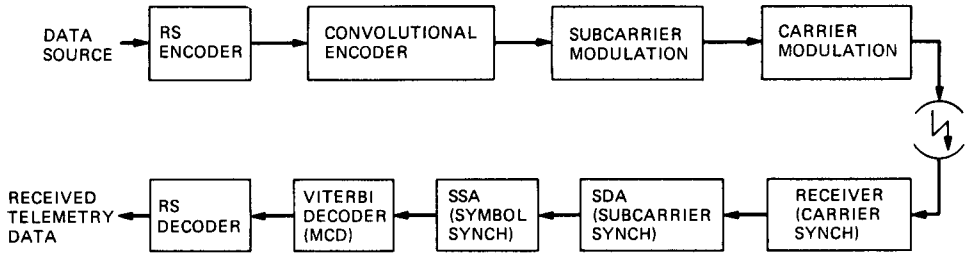


Fig. 1. Typical DSN telemetry chain.

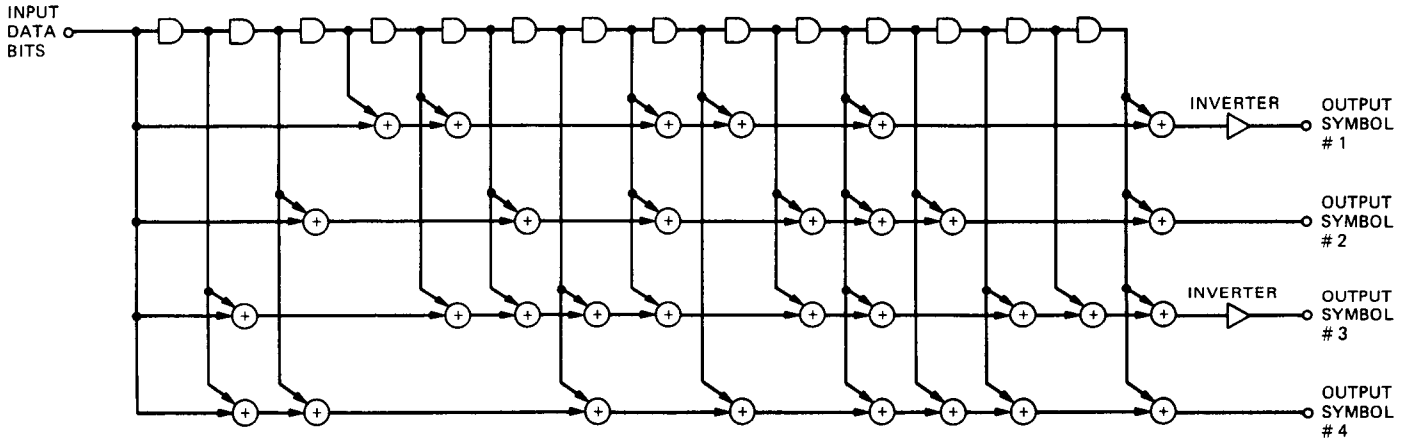


Fig. 2. A (15,¼) convolutional encoder for Galileo.

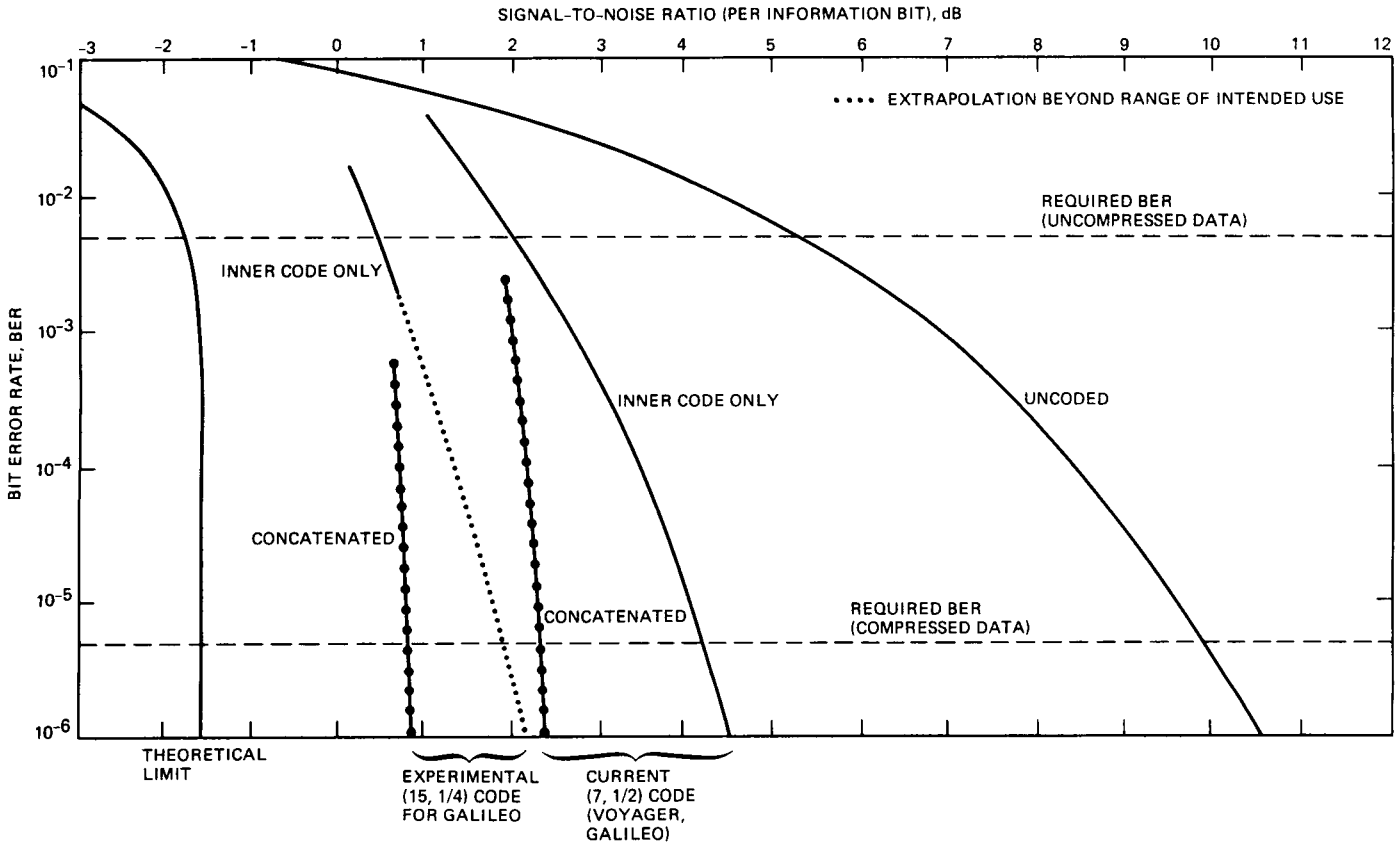


Fig. 3. Code performance.

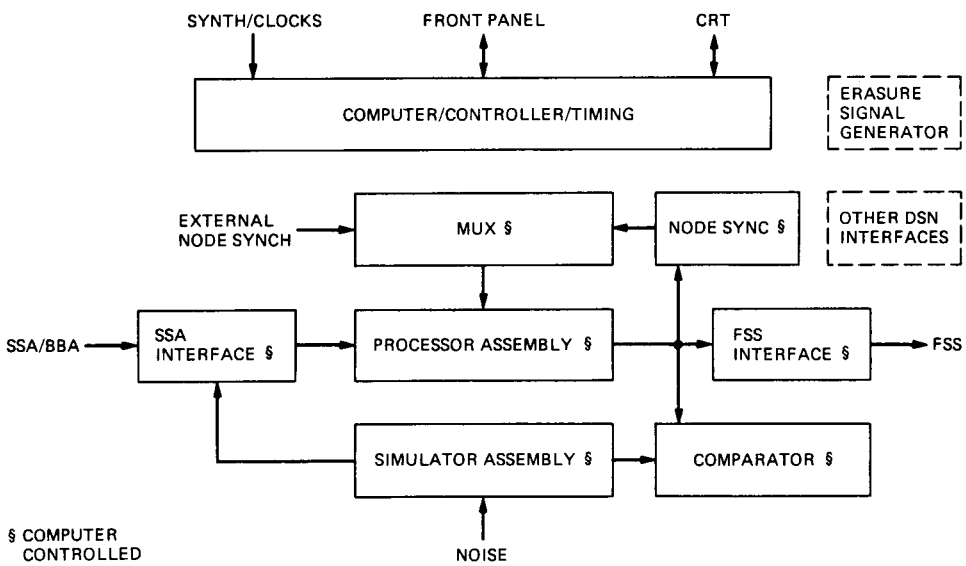


Fig. 4. Decoder functional block diagram.

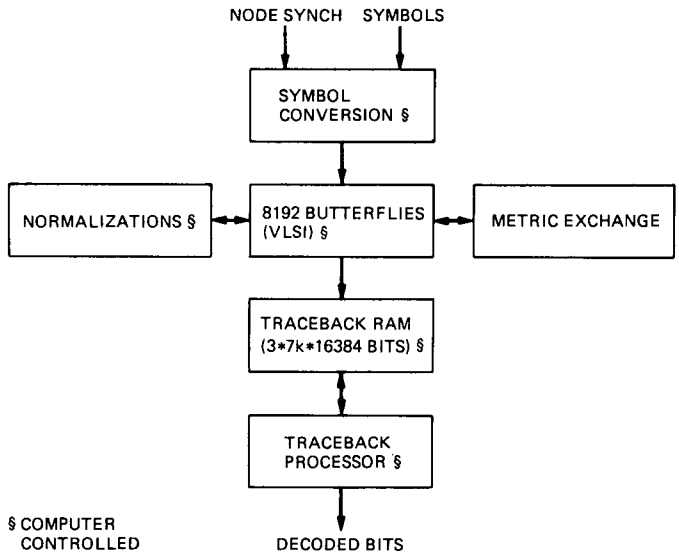


Fig. 5. Processor assembly block diagram.

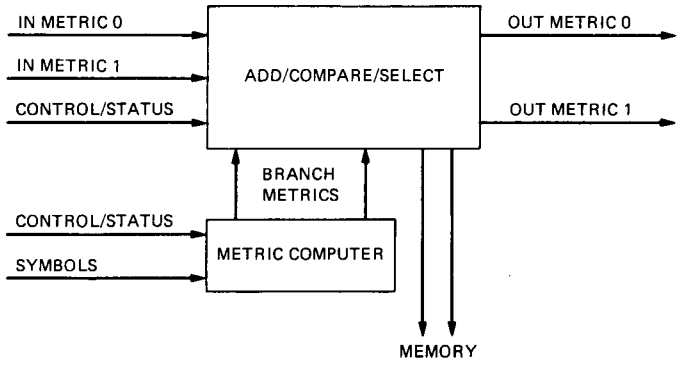


Fig. 6. Block diagram of a single butterfly.

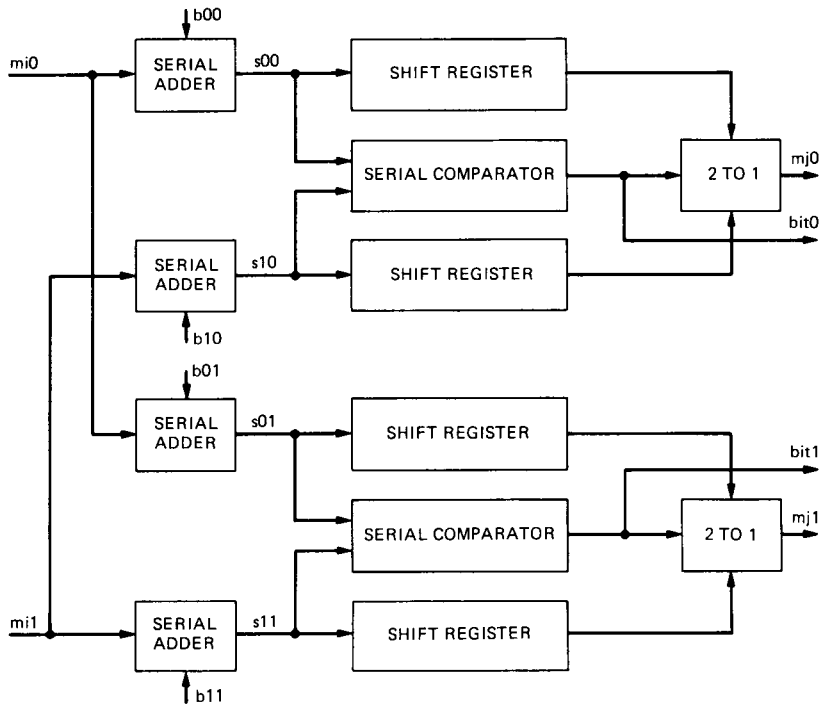


Fig. 7. Add-Compare-Select unit.

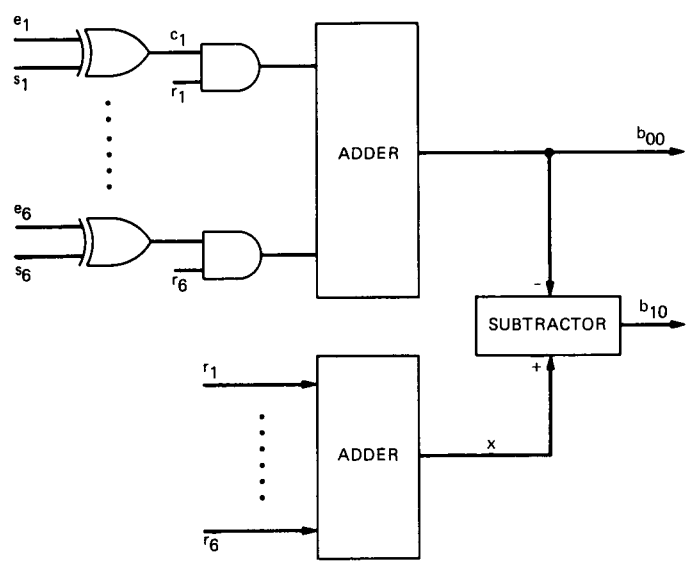


Fig. 8. Branch metric computer.

Long Decoding Runs for Galileo's Convolutional Codes

C. R. Lahmeyer and K.-M. Cheung
Communication Systems Research Section

Decoding results are described for long decoding runs of Galileo's convolutional codes. A 1-kbit/sec hardware Viterbi decoder is used for the (15, 1/4) convolutional code, and a software Viterbi decoder is used for the (7, 1/2) convolutional code. The output data of these long runs are stored in data files using a novel data compression format which can reduce file size by a factor of 100 to 1 typically. These data files can be used to replicate the long, time-consuming runs exactly and are useful to anyone who wants to analyze the burst statistics of the Viterbi decoders. The 1-kbit/sec hardware Viterbi decoder has been developed in order to demonstrate the correctness of certain algorithmic concepts for decoding Galileo's experimental (15, 1/4) code, and for long-constraint-length codes in general. The hardware decoder can be used both to search for good codes and to measure accurately the performance of known codes.

I. Introduction

Many long decoding runs of 5 to 40 Mbits have been performed for Galileo's experimental (15, 1/4) convolutional code and Galileo's standard (7, 1/2) convolutional code. A 1-kbit/sec hardware Viterbi decoder[1] is used for the (15, 1/4) convolutional code, and a software Viterbi decoder on the RTOP71 Sun-3/260 computer is used for the (7, 1/2) convolutional code. The output data from these long runs are stored in data files using a novel data compression scheme of retaining only the decoded ones and storing them in hexadecimal form. With this format a typical output compression of 100 to 1 is achieved. These data files can be used to replicate the long, time-consuming runs exactly and are useful to anyone who wants to analyze the burst statistics of the Viterbi decoders.

[1]C. R. Lahmeyer, "The 1 Kilobit per Second Viterbi Decoder," Interoffice Memorandum 331-88.3-042, Jet Propulsion Laboratory, Pasadena, California, August 19, 1988.

A 1-kbit/sec hardware Viterbi decoder was developed in the past few months in order to demonstrate the correctness of certain algorithmic concepts in the decoding of long-constraint-length convolutional codes. At present this decoder is designed to decode any convolutional code of constraint length 15 and with code rate $1/n$ as low as 1/6. Most of the recent test runs have used the (15, 1/4) convolutional code that the Galileo project has selected [1]. It has the following generator polynomials:

$$46321 = 100\ 110\ 011\ 010\ 001$$
$$51271 = 101\ 001\ 010\ 111\ 001$$
$$63667 = 110\ 011\ 110\ 110\ 111$$
$$70535 = 111\ 000\ 101\ 011\ 101$$

Since the fast hardware Viterbi decoder is at present configured to decode convolutional codes of constraint length 15 only, a software Viterbi decoder was developed in the RTOP71 Sun computer to perform long decoding runs for

Galileo's standard (7, 1/2) convolutional code. The standard code has the following generator polynomials:

$$133 = 1\ 011\ 011$$
$$171 = 1\ 111\ 001$$

The initial motives for performing the long decoding runs were to facilitate the study of the proper interleaving depth for the Reed-Solomon code used by Galileo, and to develop theoretical models (e.g., the geometric burst model [2]) for the decoded output of the Viterbi decoder, from which concatenated code performance can be accurately estimated without directly simulating the entire concatenated system. These decoding run outputs represent a data base useful to anyone studying the burst nature of the output error patterns of the Viterbi decoders.

II. Description of Test Setup of the Hardware Decoder

While the (7, 1/2) convolutional code is decoded entirely in software, the long decoding runs for the (15, 1/4) convolutional code require the use of the above-mentioned hardware decoder in combination with software. The test configuration used for the decoding runs is shown in Fig. 1. The hardware decoder interfaces with a PC-compatible computer. Software in the PC generates the test data and transmits it to the hardware decoder, where the most computationally intensive part of the decoding, called metric computation, is performed. The PC then performs the part of the decoding process called traceback, in order to complete the decoding. The PC then condenses the decoded bits into the compressed form and archives the results to hard disk.

The generation of the source data is performed by a software noise generation routine in the PC. Gaussian noise symbols are generated in software at a user-selectable noise level. The noise symbols are quantized to 8-bit sign-magnitude representation. The information content of this data is assumed to be all zeros; thus, any nonzero decoded bits represent decoding errors. This is the usual convention when running the decoder to test code performance, and it is theoretically justified because the code is linear and because it has been shown that the decoder does not favor zeros in any way.

III. Data Representation Scheme

A typical method of representing decoded output is to print ASCII 1's and 0's for all the decoded bits, but such an approach would produce a 5-Mbyte DOS file for a 5-Mbit decoding run. Therefore a compact representation scheme was developed which preserves all of the information about the decoder output but reduces file size by a factor of 100 to 1 typically. This scheme relies on the fact that the information content is all 0's, and thus the vast majority of the decoded bits will be 0's, with only a few 1's representing decoding errors. Using a scheme somewhat like spacecraft image compression, only the "changes," or in this case the error bursts, are printed. These are represented in hexadecimal notation.

Figure 2 is a sample printout of a decoding run at a 0.45-dB signal-to-noise ratio (E_b/N_0). The first column is a decimal representation of the number of bits between the start of this burst and the start of the previous one. The first burst started at bit number 0 and the second burst starts 381 bits later. Following this is a hexadecimal representation of the error burst itself. For example, 9100 represents a burst of three error bits given as 1001000100000000 in binary. The definition of the end of a burst is a string of at least 16 bits which are all 0's. Four 0's are printed at the end of each burst to act as a delimiter between bursts and to signify the 16 zero bits. This definition is somewhat loose in that some of the last bits of the printed burst can also be 0's, but no information is lost in any case. All decoding errors will ultimately be listed once each. The line with −1 signifies the end of the decoding run. Thereafter follow some statistics about the entire run. Most are self-explanatory, with Pb representing the bit error rate and Ps signifying the symbol error rate, i.e., the fraction of Reed-Solomon symbols (8 decoded bits) that are corrupted. The size of the entire file "bursts.45" is 46 kbytes.

Table 1 lists all the files accumulated so far for the Galileo code given above. Represented here are several runs of 5 Mbits and a few at 20 Mbits or more. The filename indicates the noise level used in that run, e.g., "bursts.45" signifies a run with $E_b/N_0 = 0.45$ dB.

Table 2 lists files recently generated by software decoding of the NASA standard (7, 1/2) code used by Voyager, Galileo, and other missions. In the filename, "7" signifies the constraint length and the next digits give the noise level. For example, "nburst7.1.4" signifies a run with the (7, 1/2) code at $E_b/N_0 = 1.4$ dB.

IV. Conclusions

A library of decoded output data from long decoding runs of Galileo's convolutional codes has been started. Some early runs in this collection are listed here, and it is anticipated that many more runs with different codes and sample sizes will be performed in the future. Any of the files referenced in Tables 1 and 2 can be made available to interested users on request. Useful applications of this work have already been obtained in the study of how the experimental Galileo convolutional code performs when concatenated with the 8-bit (255, 223) Reed-Solomon code [3].

References

[1] S. Dolinar, "A New Code for Galileo," *TDA Progress Report 42-93*, vol. January–March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 83–96, May 15, 1988.

[2] R. Miller, L. Deutsch, and S. Butman, *On the Error Statistics of Viterbi Decoding and the Performance of Concatenated Codes*, JPL Publication 81-9, September 1, 1981.

[3] K. Cheung and S. Dolinar, "Performance of Galileo's Concatenated Codes with Nonideal Interleaving," *TDA Progress Report 42-95*, this issue.

Table 1. Decoding runs to date with the (15, 1/4) convolutional code

Filename	Total bits decoded
bursts.0	5 Mbits
bursts.12	5 Mbits
bursts.22	5 Mbits
bursts.3	600 kbits
bursts.32	5 Mbits
bursts.42	5 Mbits
bursts.45	5 Mbits
bursts.5	22 Mbits
bursts.6	20 Mbits
bursts.7	40 Mbits

Table 2. Software decoding runs for the (7, 1/2) convolutional code

Filename	Total bits decoded
nburst7.1.4	10 Mbits
nburst7.1.45	5 Mbits
nburst7.1.5	5 Mbits
nburst7.1.55	10 Mbits
nburst7.1.6	5 Mbits
nburst7.1.65	10 Mbits
nburst7.1.7	5 Mbits
nburst7.1.75	10 Mbits
nburst7.1.8	5 Mbits
nburst7.1.85	10 Mbits
nburst7.1.9	40 Mbits
nburst7.2.0	40 Mbits

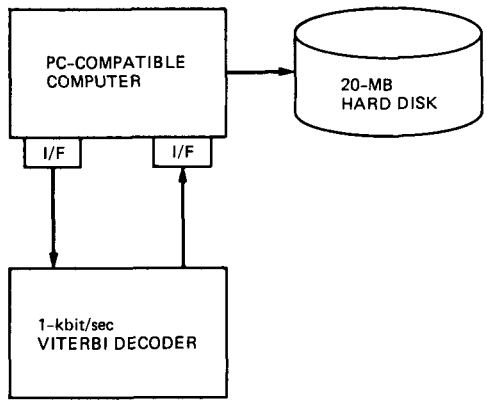


Fig. 1. Configuration for decoding with 1-kbit/sec Viterbi decoder.

```
    0 9100 0000
  381 81aa 5c0b 3a46 8000 0000
11796 bd3f 609a 1d00 0000
 4643 ef3b fd68 0000
 1490 f371 d000 0000
 4531 8afa 5eef 03ec 8000 0000
 4946 aef8 3b8e 4400 0000
 1898 819f e4fa 8c13 5d4c eb20 0000
 9206 8b62 6221 18f9 f400 0000
  612 d38e 8000 0000
 2621 f4ed e3a0 0000
 1556 f5c6 1c00 0000
                    .
                    .
                    .
                    .
                    .
                    .
 5176 c10b dcff 1f14 f258 0ab9 557b 0341 0000
 1467 b630 0000
 6557 f99e 3a00 0000
 6786 81b7 843f 086b e3a0 0000
  185 c100 d9b0 4618 0000
 4962 8a00 0000
  172 c000 0000
 2469 adb2 b389 1d64 e000 0000
 1395 850d 5e96 e755 125b 04fd cec4 9800 0000
 3468 e800 0000
 3715 c727 2720 0000
 1379 8800 0000
  651 9302 a64f 6c00 0000
  405 cc64 c1cc 4b21 f515 b200 0000
   -1

    5000040 Total bits decoded
       1560 Total bursts detected

          bits =  5000040   biterrs =   30964    Pb = 3.425e-003
          syms =   625005   symerrs =    8732    Ps = 8.343e-003
          saturation values = 0

Data recorded at  Eb/No =   0.45  db
```

Fig. 2. Sample decoding run output (from file "bursts.45").

Performance of Galileo's Concatenated Codes With Nonideal Interleaving

K.-M. Cheung and S. J. Dolinar, Jr.
Communications Systems Research Section

The Galileo spacecraft employs concatenated coding schemes with Reed-Solomon interleaving depth 2. This article compares the bit error rate (BER) performance of Galileo's concatenated codes, assuming different interleaving depths (including infinite interleaving depth). It is observed that Galileo's depth 2 interleaving, when used with the experimental (15, 1/4) code, requires about 0.4 dB to 0.5 dB additional signal-to-noise ratio to achieve the same BER performance as the concatenated code with ideal interleaving. When used with the standard (7, 1/2) code, depth 2 interleaving requires about 0.2 dB more signal-to-noise ratio than ideal interleaving.

I. Background

The Galileo spacecraft employs a communication system which uses either a (7, 1/2) convolutional code or a (15, 1/4) convolutional code as the inner code, and a (255, 223) Reed-Solomon code as the outer code. By using soft, maximum-likelihood decoding on the received symbols, the convolutional codes perform well at low signal-to-noise ratios. However, maximum-likelihood decoding of convolutional codes creates bursty errors. An interleaver is placed between the convolutional code and the Reed-Solomon code to randomize the bursty errors before they are fed to the Reed-Solomon decoder.

Concerns were initially expressed last summer[1,2] and recently repeated[3] [1] about the adequacy of Galileo's interleaving depth for the constraint length 15 code, even when it was

[1] S. Dolinar, "Alternative Code Considerations for Galileo," JPL Interoffice Memorandum 331-87.2-308 (Appendix), (internal document), Jet Propulsion Laboratory, Pasadena, California, July 1, 1987.

[2] S. Lushbough, "Galileo Convolutional Coder Meeting with Project, 31 July 1987," JPL Interoffice Memorandum 313/5-181-SL:jlc, (internal document), Jet Propulsion Laboratory, Pasadena, California, August 3, 1987.

[3] L. Swanson, "Interleaving Depths for Reed-Solomon Decoders," JPL Interoffice Memorandum 331-88.2-042, (internal document), Jet Propulsion Laboratory, Pasadena, California, July 13, 1988.

first mistakenly assumed that Galileo's interleaver was the same as Voyager's. Depth 4 interleaving was selected for Voyager to sufficiently randomize the error bursts created by the (7, 1/2) convolutional decoder. Performance degradation for the (7, 1/2) code with depth 4 interleaving is insignificant (less than 0.1 dB) relative to ideal interleaving at bit error rates between 10^{-5} and 10^{-6}. However, the error bursts from the (15, 1/4) decoder are about twice as long (on the average) as the bursts from the (7, 1/2) decoder, and thus the longer-constraint-length code would seem to require about double the interleaving depth. Instead, Galileo's actual interleaving depth is only half of Voyager's, and this can potentially cause significant concatenated system performance degradation for both of Galileo's codes relative to theoretical predictions based on ideal interleaving.

Previous studies of the effects of interleaving depth on concatenated system performance included some test data for the (7, 1/2) code[4] but no in-depth analyses that would allow extrapolation to the case of the (15, 1/4) code. Direct simulation tests of concatenated system performance using the (15, 1/4) code were completely unfeasible because of the huge amount of data that would have to be collected to verify bit error rates (BERs) in the 10^{-5} to 10^{-6} range, and because of the slowness of the software Viterbi decoder simulation (about 30 hours of CPU time on a Sun-3/260 computer per 100,000 decoded bits for the (15, 1/4) code).

Recently the completion of C. R. Lahmeyer's 1-kbit/sec (currently constrained to run at about 0.1 kbit/sec—still a hundredfold increase in speed relative to last summer's software simulation) hardware Viterbi decoder[5] has allowed us to do some long decoding runs not previously feasible for the (15, 1/4) code. With the advent of this hardware decoder, the following research tools have been developed:

(1) Long decoding runs (several megabits) for the (15, 1/4) convolutional code were performed on the hardware decoder, and the error bursts are stored in data files conforming to the data compression format described in recent memos.[6,7] These data files can be used to replicate the long, time-consuming runs exactly and are useful to anyone who wants to analyze the burst statistics of the Viterbi decoder.

(2) Similar long decoding runs were performed for the (7, 1/2) convolutional code using the software simulation, and the error bursts from those runs are also saved in the compressed format.

(3) Simulation software was developed which reads the compressed burst data obtained from the long decoding runs and simulates the operation of the entire concatenated coding system with different interleaving depths.

II. Performance Results

Simulated BERs of concatenated coding systems with various interleaving depths are given in Figs. 1 and 2. Figure 1 shows the performance of the concatenated system using the (15, 1/4) convolutional code as the inner code, and Fig. 2 shows the performance of the concatenated system using the (7, 1/2) convolutional code. The BERs of the concatenated coding systems with finite interleaving depths are compared to the BERs with ideal interleaving (infinite interleaving depth). In both figures, the concatenated code BER is shown as a function of two bit-energy-to-spectral-noise-density ratios, E_b/N_0:

(1) for the convolutional code alone and

(2) for the overall concatenated system.

The difference between the two E_b/N_0 scales is the overhead of 0.58 dB accounting for the redundancy of the outer (255, 223) Reed-Solomon code.

The data points plotted in Figs. 1 and 2 were obtained from a series of long decoding runs varying in length from 5 million to 40 million decoded bits. Smooth curves were fitted through the data points corresponding to BERs greater than 10^{-5}. The 1σ statistical uncertainty in the data points at BERs lower than 10^{-5} is more than 100% for the cases of finite interleaving depth. The corresponding uncertainty in the data points at BERs higher than 10^{-5} ranges from about 5% to about 100% of the simulated BER. The 1σ uncertainty in the data points for ideal interleaving is between 10% and 30% of the BER.

An example illustrates the difficulty of obtaining accurate simulated concatenated system performance at low BERs. The rightmost data point in Fig. 1 (interleaving depth 2, convolutional code E_b/N_0 = 0.7 dB) required about 120 hours (5 days) of running time on the hardware decoder to decode 40 Mbits. The same decoding run would have consumed

[4] F. H. J. Taylor, "Project Galileo Orbiter/Deep Space Network Communications Design Document," Project Document 625-257, Jet Propulsion Laboratory, Pasadena, California, pp. 5–12, January 15, 1981.

[5] C. Lahmeyer, "1 Kilobit per Second Viterbi Decoder," JPL Interoffice Memorandum 331-88.3-042, (internal document), Jet Propulsion Laboratory, Pasadena, California, August 19, 1988.

[6] C. Lahmeyer and K. Cheung, "Long Decoding Runs for the (15, 1/4) Convolutional Code," JPL Interoffice Memorandum 331-88.3-040, (internal document), Jet Propulsion Laboratory, Pasadena, California, July 27, 1988.

[7] K. Cheung, "More Long Decoding Runs of the Convolutional Codes (including the (15, 1/4) convolutional code and the (7, 1/2) convolutional code)," JPL Interoffice Memorandum 331-88.2-048, (internal document), Jet Propulsion Laboratory, Pasadena, California, August 1, 1988.

1.4 *years* of CPU time on the software decoder. Furthermore, the entire 40-Mbit run produced only three observed codeword errors, and the 1σ statistical uncertainty in the simulated BER is around 100%.

III. Performance Comparison

Tables 1 and 2 show the minimum convolutional code E_b/N_0 to achieve Galileo's concatenated and unconcatenated system performance requirements assuming ideal (infinite-depth) interleaving and depth 2 interleaving, respectively. Table 3 shows the performance degradation caused by depth 2 interleaving relative to ideal interleaving for Galileo's two alternative convolutional codes. Galileo's experimental (15, 1/4) code requires about 0.4 dB to 0.5 dB additional signal-to-noise ratio to overcome the insufficiencies of depth 2 interleaving and achieve concatenated code BERs between 10^{-5} and 10^{-6}. Galileo's standard (7, 1/2) code is hurt less by depth 2 interleaving, but still suffers about 0.2 dB degradation. The relative performance advantage of the (15, 1/4) code over the (7, 1/2) code is reduced by about 0.2 dB to 0.3 dB from the amount predicted in earlier studies (e.g., [1]) based on ideal interleaving. With depth 2 interleaving, concatenated system performance will be improved by only about 1.2 dB when the (15, 1/4) code is substituted for the (7, 1/2) code. The corresponding improvement for an unconcatenated system or for a concatenated system with ideal interleaving is between 1.4 dB and 1.5 dB.

IV. Conclusions and Recommendations

Galileo is unfortunately stuck with depth 2 interleaving, and so the immediate consequence of our simulations is simply to quantify the amount of expected degradation for Galileo's concatenated codes. However, future missions should select interleaving schemes that produce minimal degradation. The required interleaving depth increases roughly in proportion to the constraint length of the inner convolutional code. For example, interleaving depth 8 appears sufficient to keep the degradation under 0.1 dB for the (15, 1/4) code, as does interleaving depth 4 for the (7, 1/2) code.

As alternatives to simply increasing the interleaving depth of conventional block interleaving schemes, new techniques to combat bursty errors, such as convolutional interleaving, helical interleaving, and burst forecasting, should also be investigated. These techniques are superior to conventional block interleaving schemes. Also, a new Reed-Solomon decoder which can correct both errors and erasures is being developed in the Communications Systems Research Section at JPL. It is expected that the performance of concatenated systems will be substantially improved by the use of error-forecasting techniques together with erasure-correcting Reed-Solomon decoders. We propose to investigate the possibility of developing better interleaving schemes for future deep space missions and to analyze the performance of concatenated coding systems using these new interleaving schemes.

Even though the hardware Viterbi decoder has allowed us to simulate many million decoded bits at a time, the data is still insufficient to accurately simulate concatenated code performance at BERs less than about 10^{-5}. The amount of data required for an accurate estimate increases in proportion to the interleaving depth, and so it is even more difficult to simulate directly the performance of deeply interleaved schemes than it was for Galileo's interleaving depth 2. Hence, notwithstanding the recent advance in decoding speed, it is still important to develop theoretical models for the decoded output of the Viterbi decoder, from which concatenated code performance can be accurately estimated without directly simulating the entire concatenated system. The geometric burst model of [2] should be reexamined for applicability tc long-constraint-length codes, and new models need to be developed for estimating Reed-Solomon code performance based on the theoretical model for the Viterbi decoder output.

References

[1] S. Dolinar, "A New Code for Galileo," *TDA Progress Report 42-93*, January–March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 83–96, May 15, 1988.

[2] R. L. Miller, L. J. Deutsch, and S. A. Butman, *On the Error Statistics of Viterbi Decoding and the Performance of Concatenated Codes*, JPL Publication 81-9, Jet Propulsion Laboratory, Pasadena, California, September 1, 1981.

Table 1. Minimum E_b/N_0 to achieve concatenated and unconcatenated system performance requirements under ideal interleaving assumption

	(7, 1/2) code	(15, 1/4) code	Difference
Unconcatenated BER = 5×10^{-3}	2.02 dB	0.52 dB	1.50 dB
Concatenated BER = 10^{-6}	1.79 dB	0.33 dB	1.46 dB
Concatenated BER = 10^{-5}	1.70 dB	0.27 dB	1.43 dB

Table 2. Minimum E_b/N_0 to achieve concatenated and unconcatenated system performance requirements for interleaving depth 2

	(7, 1/2) code	(15, 1/4) code	Difference
Unconcatenated BER = 5×10^{-3}	2.02 dB	0.52 dB	1.50 dB
Concatenated BER = 10^{-6}	2.01 dB	0.83 dB	1.18 dB
Concatenated BER = 10^{-5}	1.90 dB	0.69 dB	1.21 dB

Table 3. Concatenated system performance degradation for interleaving depth 2 versus ideal interleaving

	(7, 1/2) code	(15, 1/4) code	Difference
Concatenated BER = 10^{-6}	0.22 dB	0.50 dB	0.28 dB
Concatenated BER = 10^{-5}	0.20 dB	0.42 dB	0.22 dB

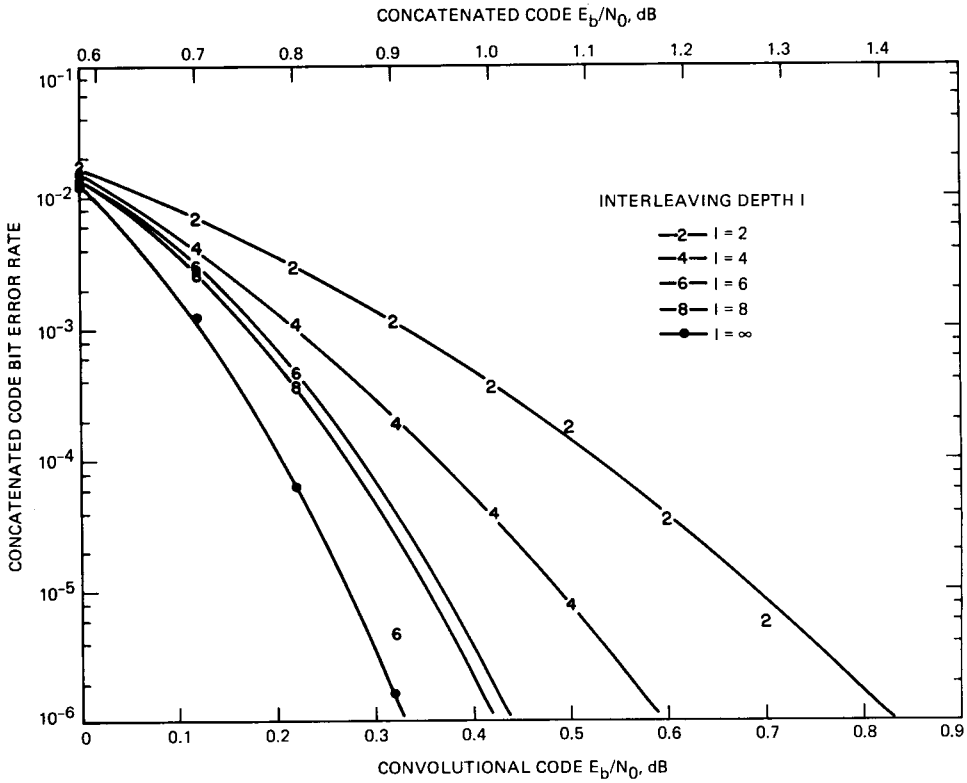


Fig. 1. Concatenated code performance for Galileo's experimental (15, 1/4) inner code with nonideal interleaving.

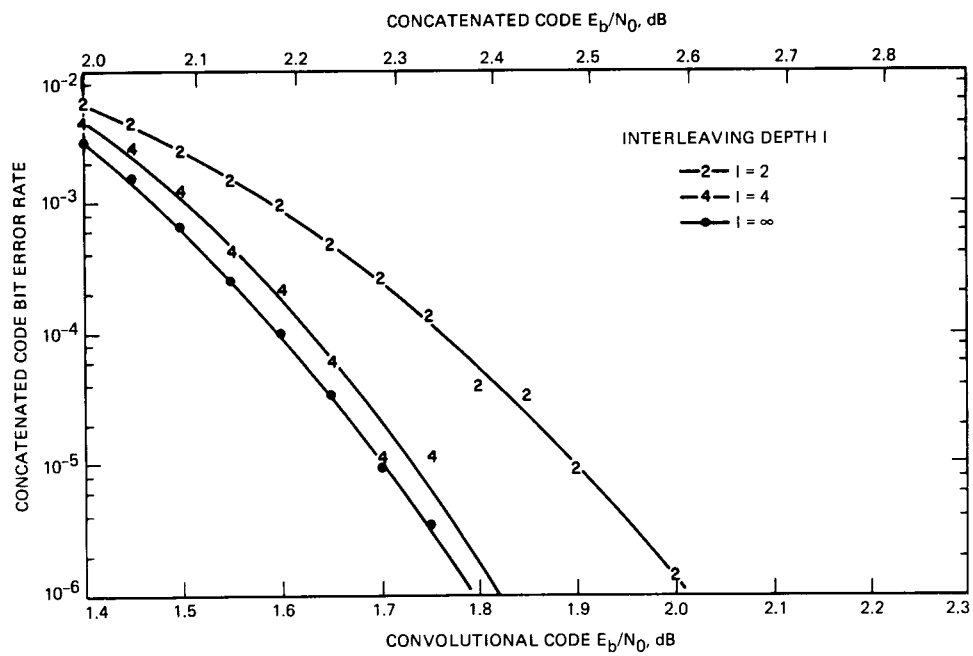


Fig. 2. Concatenated code performance for Galileo's standard (7, 1/2) inner code with nonideal interleaving.

The Decoding of Reed-Solomon Codes

R. J. McEliece
Communications Systems Research Section
Electrical Engineering Department
California Institute of Technology

Reed-Solomon (RS) codes form an important part of the high-rate downlink telemetry system for the Magellan mission, and the RS decoding function for this project will be done by the DSN. Although the basic idea behind all Reed-Solomon decoding algorithms was developed by Berlekamp in 1968, there are dozens of variants of Berlekamp's algorithm in current use. This paper attempts to restore order by presenting a mathematical theory which explains the working of almost all known RS decoding algorithms. The key innovation that makes this possible is the unified approach to the solution of the key equation, which simultaneously describes the Berlekamp, Berlekamp-Massey, Euclid, and continued fractions approaches. Additionally, a detailed analysis is made of what can happen to a generic RS decoding algorithm when the number of errors and erasures exceeds the code's designed correction capability, and it is shown that while most published algorithms do not detect as many of these error-erasure patterns as possible, by making a small change in the algorithms, this problem can be overcome.

I. Decoding Reed-Solomon Codes

In this article we will give a general definition of Reed-Solomon codes, state the abstract Reed-Solomon decoding problem, describe the two main classes of decoding algorithms (time- and frequency-domain decoders), and then give three theorems. Theorem 1 explains why the RS decoding algorithms work, and Theorems 2 and 3 delineate exactly what happens if the number of errors and erasures exceeds the codes' designed correction capability. The article concludes with three appendices, which give the mathematical background needed for the proofs of the theorems presented.

Let F be a field which contains a primitive nth root of unity α. (We assume that the characteristic of the field does not divide n.) If L and r are fixed integers between 0 and n, the set of all codewords (vectors) $\mathbf{C} = (C_0, \ldots, C_{n-1})$ over F such that

$$\sum_{i=0}^{n-1} C_i \alpha^{ij} = 0 \qquad j = L, L+1, \ldots, L+r-1 \qquad (1)$$

is called a Reed-Solomon code. The parameters n, r, and L are called the *length*, *redundancy*, and *offset* of the code. The

parameters $k = n - r$ is called the code's dimension. (Commonly $L = 0$, 1, or $(n - r + 1)/2$.) The polynomial $g(x)$, defined by

$$g(x) = \prod_{j=L}^{L+r-1} (x - \alpha^j) \tag{2}$$

is called the code's *generator polynomial*. Note that $\mathbf{C}$ is a codeword if and only if the generating function for $\mathbf{C}$, viz., $C(x) = C_0 + C_1 x + \cdots + C_{n-1}x^{n-1}$ is a multiple of $g(x)$. If n is odd and r is even, and $L = (n - r + 1)/2$, then the roots of $g(x)$ come in reciprocal pairs (α^j, α^{-j}), for $j = L, \ldots, L + (r/2) - 1$, and $g(x)$ is a "palindrome."

The *basic metric property* of an RS code is that any two codewords must differ in at least $r + 1$ positions. Thus if $d_{\mathbf{H}}(\mathbf{C}, \mathbf{C}')$ denotes the Hamming distance between the codewords $\mathbf{C}$ and $\mathbf{C}'$, and $\mathbf{C} \neq \mathbf{C}'$, then it follows that

$$d_{\mathbf{H}}(\mathbf{C}, \mathbf{C}') \geqslant r + 1 \tag{3}$$

The *basic combinatorial property* of an RS code is that given any subset $I \subseteq \{0, 1, \ldots, n - 1\}$ of at most k coordinate positions, and an arbitrary set $\{\alpha_i : i \in I\}$ of elements from F, there exists an RS codeword $\mathbf{C}$ such that $C_i = \alpha_i$ for all $i \in I$. (Proofs of these basic properties can be found in [2], Section 7.3.)

Suppose we transmit a codeword $\mathbf{C}$ over a channel, which can, on occasion, change any symbol from F into any other, and which in addition can "erase" any symbol, i.e., make it completely unintelligible. To model erasures, we introduce an "erasure symbol" $*$ and add it to $F : \bar{F} = F \cup \{*\}$. Thus we send a codeword, and receive a vector $\mathbf{R} = (R_0, R_1, \ldots, R_{n-1})$ from $\bar{F}^n$. The RS *decoding problem*, ideally, would be this: given $\mathbf{R} \in \bar{F}^n$, find the nearest RS codeword. However, that proves to be too hard, and we must be satisfied with the solution to an easier but closely related problem.

To state the decoding problem precisely, we must define a distance between vectors over $\bar{F}$, the RS *decoding distance*. If $\mathbf{V} = (V_0, \ldots, V_{n-1})$ and $\mathbf{V}' = (V_0', \ldots, V_{n-1}')$ are vectors with components in $\bar{F}$, we define

$$d_{\mathbf{RS}}(\mathbf{V}, \mathbf{V}') = \sum_{i=0}^{n-1} d_{\mathbf{RS}}(V_i, V_i') \tag{4}$$

where if x and y are elements of $\bar{F}$,

$$d_{\mathbf{RS}}(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \text{ and either } x \text{ or } y \text{ is } * \\ 2 & \text{if } x \neq y \text{ but neither } x \text{ nor } y \text{ is } * \end{cases} \tag{5}$$

One way to think about this metric is shown in Fig. 1, for $F = GF(3)$. The elements of $\bar{F}$ are the vertices of a graph, with every element of F connected to $*$ by an edge. Then $d_{\mathbf{RS}}(x, y)$ is just the distance between x and y in the graph. Note that if $\mathbf{V}$ and $\mathbf{V}'$ are vectors with components in F (i.e., with no $*$s) we have

$$d_{\mathbf{RS}}(\mathbf{V}, \mathbf{V}') = 2 d_{\mathbf{H}}(\mathbf{V}, \mathbf{V}') \tag{6}$$

The *decoding problem* for an RS code of redundancy r can now be stated. Given an arbitrary vector $\mathbf{R} = (R_0, \ldots, R_{n-1})$ with components from $\bar{F}$, find all RS codewords $\mathbf{C}$ such that

$$d_{\mathbf{RS}}(\mathbf{C}, \mathbf{R}) \leqslant r \tag{7}$$

First, note that there can be at most one codeword $\mathbf{C}$ such that Eq. (7) holds. This is because if $d_{\mathbf{RS}}(\mathbf{C}, \mathbf{R}) \leqslant r$ and $d_{\mathbf{RS}}(\mathbf{C}', \mathbf{R}) \leqslant r$, then by the triangle inequality

$$d_{\mathbf{RS}}(\mathbf{C}, \mathbf{C}') \leqslant d_{\mathbf{RS}}(\mathbf{C}, \mathbf{R}) + d_{\mathbf{RS}}(\mathbf{R}, \mathbf{C}') \leqslant 2r \tag{8}$$

which implies by Eq. (5) that $d_{\mathbf{H}}(\mathbf{C}, \mathbf{C}') \leqslant r$, violating Eq. (3), unless $\mathbf{C} = \mathbf{C}'$.

We now describe an efficient algorithm, essentially due to Elwyn Berlekamp ([1], Chapter 7), for solving the decoding problem.

For a given $\mathbf{R}$, its *erasure set* I_0 is defined as

$$I_0 = \{i : R_i = *\} \qquad t_0 = |I_0| \tag{9}$$

(The notation $|S|$ denotes the number of elements in the set S.) The decoder's first step is to calculate the *erasure locator polynomial* $\sigma_0(x)$, defined by

$$\sigma_0(x) = \prod_{i \in I_0} (1 - \alpha^i x) \tag{10}$$

(If there are no erasures in $\mathbf{R}$, $\sigma_0(x)$ is defined to be 1.) If the number of erasures t_0 exceeds r, there can be no solutions to Eq. (7); in this case the decoder should simply print "too many erasures!" and stop. We will assume that $t_0 \leq r$ in the rest of the discussion of the decoding algorithm.

Once the erasure locator polynomial is calculated, the decoder replaces the *s in $\mathbf{R}$ with symbols from F. Usually these symbols are chosen to be 0s, but if the decoder has "side information" about the original values of the C_is corresponding to the erased indices $i \in I_0$, these values can be used. In any case, the result is a new vector $\mathbf{R}' = (R'_0, \ldots, R'_{n-1})$, defined by

$$R'_i = \begin{cases} R_i & \text{if } i \notin I_0 \\ Z_i & \text{if } i \in I_0 \end{cases} \tag{11}$$

where $Z_i = 0$ is the usual choice.

Next, the *syndrome* is computed, i.e., the r values

$$S_j = \sum_{i=0}^{n-1} R'_i \alpha^{ij} \qquad \text{for } j = L, L+1, \ldots, L+r-1 \tag{12}$$

which are used as coefficients in the *syndrome polynomial*

$$S(x) = S_L + S_{L+1}x + \cdots + S_{L+r-1}x^{r-1} \tag{13}$$

If erasures are present the decoder continues by calculating the *modified syndrome polynomial* $S_0(x)$, defined by

$$S_0(x) = S(x)\,\sigma_0(x) \pmod{x^r} \tag{14}$$

Now comes the key step. Define the numbers μ and ν by

$$\begin{aligned} \mu &= \lfloor (r - t_0)/2 \rfloor \\ \nu &= \lceil (r + t_0)/2 \rceil - 1 \end{aligned} \tag{15}$$

(If x is a real number, $\lfloor x \rfloor$ is the greatest integer less than or equal to x, and $\lceil x \rceil$ is the least integer greater than or equal to x.) It is an easy exercise to show that $\mu + \nu = r - 1$. The decoder now solves the $(x^r, S_0(x), \mu, \nu)$ problem, i.e., it finds the unique lowest degree pair of polynomials $\sigma_1(x)$ and $\omega(x)$ such that $\deg(\sigma_1) \leq \mu$, $\deg(\omega) \leq \nu$, and

$$\sigma_1(x)S_0(x) \equiv \omega(x) \pmod{x^r}$$

(see Appendix C). The polynomial $\sigma_1(x)$ is traditionally called the *error locator polynomial*, and $\omega(x)$ is called the *error-and-erasure evaluator polynomial*. Now the decoder multiplies $\sigma_0(x)$ and $\sigma_1(x)$, obtaining a polynomial $\sigma(x)$, called the erasure/error locator polynomial.

Once the polynomials $\sigma(x)$ and $\omega(x)$ are known, there are two essentially different ways to complete the algorithm. These are usually called the *time-domain* approach and the *frequency-domain* approach.

The *time-domain* approach can be described by the following pseudocode fragment.

```
/* Time domain fragment */
{
  if (σ0 = 0 or deg(ω) ≥ t0 + deg(σ1))
    decode = FALSE;
  else {
    count = 0;
    for (i = 0 to n - 1) {
      if (σ(α^-i) = 0 and σ'(α^-i) ≠ 0) {
        count = count + 1;
        Ei = -α^-i(L-1) ω(α^-i) / σ'(α^-i) ;
      }
      else
        Ei = 0;
    }
  if (count = deg(σ))
    decode = TRUE;
  else
    decode = FALSE;
  }
}
```

After execution, if "decode" is "TRUE," $\mathbf{C} = (C_0, \ldots, C_{n-1})$, where $C_i = R'_i - E_i$ for $i = 0, 1, \ldots, n-1$ is the unique codeword within RS distance r of $\mathbf{R}$. On the other hand, if "decode" is "FALSE," the decoder just prints the warning "no codeword within RS distance r." All early RS decoders used an algorithm much like this; such an algorithm is described as a "hybrid decoder" in Figure 9.2 in Blahut [2].

The frequency-domain approach can be described by the following pseudocode fragment. (In this listing, d denotes the degree of $\sigma(x)$.)

```
/* Frequency Domain Fragment */
{
   if (σ0 = 0)
      decode = FALSE;
   else {
      decode = TRUE;
      for (j = L + r to n + L + d - 1)
         S_j = -(1/σ0) Σ_{k=1}^{d} σ_k S_{j-k};
      for (j = n + L to n + L + d - 1)
         if (S_j ≠ S_{j-n})  {
            decode = FALSE;
            break;
         }
   }
   if (decode = TRUE)
      for (i = 0 to n - 1)
         E_i = α^{-Li} · (1/n) Σ_{j=0}^{n-1} S_{L+j} α^{-ij};
}
```

The decoder now finishes exactly as the time-domain decoder did. The "frequency-domain decoders" described in [2, Figure 9.2] and the decoder described in [5] follow this general description. (The "time-domain decoder" described in Figure 9.7 in [2] is a rare example of an RS decoding algorithm which is apparently not closely related to the descriptions in this section. See Whiting [6] for a survey of Reed-Solomon decoding algorithms.)

In each case, the algorithm will locate the codeword within RS distance r of $\mathbf{R}$, if there is one, and will print the message "no codeword within RS distance r" if there is not. The following theorem explains why.

Theorem 1. There is a codeword within RS distance r of $\mathbf{R}$ if and only if the following three conditions are satisfied:

(A) $\deg(\omega(x)) < t_0 + \deg(\sigma_1(x))$

(B) $\sigma_1(0) \neq 0$

(C) $\sigma_0(x)\,\sigma_1(x) \mid (1 - x^n)$

Proof: First, we suppose there is a codeword $\mathbf{C} = (C_0, \ldots, C_{n-1})$ within RS distance r of $\mathbf{R}$. We will show that conditions (A), (B), and (C) are satisfied. To do this, we define I_0, $\sigma_0(x)$, and $\mathbf{R}'$ as in Eq. (11) (the erasure fills Z_i can be arbitrary). Next, we define the *error set* I_1 and *error locator polynomial* $\sigma_1(x)$ as

$$I_1 = \{i \notin I_0 : R_i \neq C_i\} \tag{16}$$

$$\sigma_1(x) = \prod_{i \in I_1} (1 - \alpha^i x) \tag{17}$$

and the *error-and-erasure pattern* as $\mathbf{E} = (E_0, \ldots, E_{n-1})$, where

$$E_i = R'_i - C_i \qquad \text{for } i = 0, \ldots, n-1 \tag{18}$$

Finally we define the *error-and-erasure set* I and the *error-and-erasure locator polynomial* $\sigma(x)$ by

$$I = I_0 \cup I_1 \tag{19}$$

$$\sigma(x) = \prod_{i \in I} (1 - \alpha^i x) \tag{20}$$

It follows from Eqs. (12), (18), and (1) that the syndromes S_j satisfy

$$S_j = \sum_{i=0}^{n-1} E_i \alpha^{ij} \qquad \text{for } j = L, \ldots, L + r - 1 \tag{21}$$

which implies that

$$S_{j+L} = \sum_{i=0}^{n-1} (E_i \alpha^{iL})\, \alpha^{ij} \qquad \text{for } j = 0, \ldots, r-1 \tag{22}$$

Thus $S_L, S_{L+1}, \ldots, S_{L+r-1}$ are the first r components of the DFT $\hat{\mathbf{V}} = (\hat{V}_0, \ldots, \hat{V}_{n-1})$ of the "twisted error pattern" $\mathbf{V} = (V_0, V_1, \ldots, V_{n-1})$, defined by

$$V_i = E_i \alpha^{iL} \qquad \text{for } i = 0, \ldots, n-1 \tag{23}$$

It follows then from Eqs. (13) and (22) that if we define $\hat{V}(x) = \hat{V}_0 + \hat{V}_1 x + \cdots + \hat{V}_{n-1} x^{n-1}$, then $\hat{V}(x) \equiv S(x) \pmod{x^r}$, and indeed, if we define $\hat{V}_0(x) = \sigma_0(x)\, \hat{V}(x) \bmod x^r$, that

$$\hat{V}_0(x) = S_0(x) \tag{24}$$

where $S_0(x)$ is defined in Eq. (14). If we now define the *error-and-erasure evaluator polynomial* as

$$\omega(x) = \sum_{i \in I} V_i \sigma^i(x) \tag{25}$$

where $\sigma^i(x) = \sigma(x)/(1 - \alpha^i x)$, (compare this to Eq. (B-5) in Appendix B), it follows from Theorem B-6 in Appendix B that

$$\sigma(x)\,\widehat{V}(x) = \omega(x)\,(1 - x^n) \tag{26}$$

and so, since $\sigma(x) = \sigma_0(x)\,\sigma_1(x)$ and $\widehat{V}_0(x) = \sigma_0(x)\,\widehat{V}(x)$,

$$\sigma_1(x)\,\widehat{V}_0(x) \equiv \omega(x) \pmod{x^r} \tag{27}$$

Furthermore, since **C** is assumed to have RS distance r or less from **R**, it follows that $t_0 + 2\deg(\sigma_1) \leqslant r$, which in turn implies $\deg(\sigma_1) \leqslant \mu$, and $\deg(\omega) < \deg(\sigma) = t_0 + \deg(\sigma_1) \leqslant \nu$, where μ and ν are defined in Eq. (15). Furthermore, σ_1 and ω are relatively prime, since by Lemma B-2 in Appendix B for each $i \in I_1$, $\omega(\alpha^{-i}) \neq 0$. Therefore (σ_1, ω) is the solution to the $(x^r, \widehat{V}_0(x), \mu, \nu)$ problem, which by Eq. (24) is the same as the $(x^r, S_0(x), \mu, \nu)$ problem. Thus the polynomials produced by the decoding algorithm must be the error locator polynomial and the error-and-erasure evaluator polynomial, and these polynomials satisfy conditions (A), (B), and (C): Equation (25) implies (A); Equation (10) implies (B); Equation (20) implies (C).

To complete the proof, we suppose that conditions (A), (B), and (C) are satisfied. We will show that this implies that there is a codeword within RS distance r of **R**. To do this we define $\sigma(x) = \sigma_0(x)\,\sigma_1(x)$; note that condition (A) says that $\deg(\omega) < \deg(\sigma)$, and condition (C) says that $\sigma(x) | (1 - x^n)$. Hence by Theorem B-5 in Appendix B, there exists a vector $\mathbf{V} = (V_0, V_1, \ldots, V_{n-1})$ and a support set I for **V** such that

$$\sigma(x) = \lambda \sigma_I(x) \tag{28}$$

$$\omega(x) = \lambda \omega_{\mathbf{V},I}(x) \tag{29}$$

We claim now that the vector $\mathbf{C} = (C_0, \ldots, C_{n-1})$, defined by

$$C_i = R_i' - V_i \alpha^{-iL} \qquad \text{for } i = 0, \ldots, n-1 \tag{30}$$

is a codeword within RS distance r of **R**. First we show that $d_{\mathbf{RS}}(\mathbf{C}, \mathbf{R}) \leqslant r$. This is because **R** has t_0 erasure symbols, and apart from these, differs from **C** only in those indices i for which $V_i \neq 0$, i.e., $\sigma_1(\alpha^{-i}) = 0$. But by condition (C), σ_1 has exactly $\deg(\sigma_1)$ roots in $\{1, \alpha, \ldots, \alpha^{n-1}\}$, and $\deg(\sigma_1) \leqslant \mu$, and so

$$d_{\mathbf{RS}}(\mathbf{C}, \mathbf{R}) = t_0 + 2\deg(\sigma_1) \leqslant t_0 + 2\mu \leqslant t_0 + 2\lfloor (r - t_0)/2 \rfloor \leqslant r \tag{31}$$

All that remains is to show that **C**, as defined in Eq. (30), is a codeword. Since σ_1 and $\omega(x)$ solve the $(x^r, S_0(x), \mu, \nu)$ problem, we know that

$$\sigma_1(x)\,S_0(x) \equiv \omega(x) \pmod{x^r} \tag{32}$$

But by Eq. (14), $S_0(x) = S(x)\,\sigma_0(x) \bmod x^r$; and since $\sigma(x) = \sigma_0(x)\,\sigma_1(x)$, by Eq. (32) we have

$$\sigma(x)\,S(x) \equiv \omega(x) \pmod{x^r} \tag{33}$$

On the other hand, by Eqs. (28) and (29), together with Theorem B-6 in Appendix B, we have

$$\sigma(x)\,\widehat{V}(x) \equiv \omega(x) \pmod{x^r} \tag{34}$$

Now $\gcd(\sigma_0(x), x^r) = 1$ (see Eq. 10), and condition (B) guarantees that $\gcd(\sigma_1(x), x^r) = 1$, and so $\gcd(\sigma(x), x^r) = 1$. Thus by Eqs. (33) and (34) we have

$$S(x) \equiv \widehat{V}(x) \pmod{x^r} \tag{35}$$

Equating coefficients of x^j for $j = 0, 1, \ldots, r-1$ on both sides of Eq. (35), we see that

$$S_{j+L} = \widehat{V}_j \qquad \text{for } j = 0, 1, \ldots, r-1 \tag{36}$$

But this implies that

$$\sum_{i=0}^{n-1} R_i' \alpha^{ij} = \sum_{i=0}^{n-1} V_i \alpha^{-iL} \alpha^{ij} \qquad \text{for } j = L, \ldots, L + r - 1 \tag{37}$$

which says that **C**, as defined by Eq. (30), is a codeword. ■

With the help of Theorem 1, we can now explain why the time-domain and frequency-domain decoders work. First, we discuss the time-domain decoder. The first line checks condition (A) and (B) of Theorem 1. The "for" loop checks condition (C), by evaluating the polynomial $\sigma(x)$ for $x = \alpha^{-i}$, for $i = 0, 1, \ldots, n-1$. Notice that there is a check for $\sigma'(\alpha^{-i}) = 0$; this is necessary, since as we will see, it is possible for $\sigma(x)$ to have a double root. The formula for E_i follows from Eq. (26) and Corollary B-7 in Appendix B, and the fact that $E_i = V_i \alpha^{-Li}$ (see Eq. 23).

Next, we consider the frequency-domain decoder. The first line checks condition (B) of Theorem 3. The first "for" loop extends the sequence $S_L, S_{L+1}, \ldots, S_{L+r-1}$ recursively, using $\sigma(x)$ as the characteristic polynomial, and the second "for" loop checks to see whether or not this extension has period n. If the sequence is not periodic, then either condition (A) or (C) must fail, by the first part of Theorem B-10 in Appendix B. On the other hand, if the sequence is periodic, then

$$\sum_{k=0}^{d} \sigma_k S_{j-k} = 0 \qquad \text{for all } j \geq L + d$$

and so if $u(x) = S_L + S_{L+1}x + \cdots$, then $\sigma(x)\,u(x)$ has degree $< d$. But $\sigma(x)\,u(x) \equiv \omega(x) \pmod{x^r}$, and so $\deg(\omega) < d$, i.e., condition (A) is satisfied. Next, since condition (B) insures that $\gcd(\sigma_1, \omega) = 1$, the second part of Theorem B-10 shows that condition (C) holds. Finally, the formula given for the error vector E_i follows from the fact that $(S_L, \ldots, S_{L+n-1})$ is the DFT $\hat{\mathbf{V}}$ of the twisted error $\mathbf{V}$ vector defined in Eq. (23). This follows from the basic Theorem B-6, which says that the components of $\hat{\mathbf{V}}$ satisfy the homogeneous difference equation whose characteristic polynomial is $\sigma(x)$.

All published RS decoding algorithms correctly locate the codeword within RS distance r of the received word, if there is one. However, almost all of these algorithms (including all algorithms in [2], [3], [5], and [6]) can fail badly when there is no such word. By this we mean that there usually exist words $\mathbf{R}$ which are not within RS distance r of any codeword, and yet which cause the decoder to output a vector $\mathbf{C}$ rather than to print the message "no RS codeword within RS distance r." One naive way to avoid this problem is simply to test any word produced by the decoder to see if it is a codeword, and then to see if it is within RS distance r of the received word. However, this method is not quite foolproof (division by zero is possible in either the time- or frequency-domain approaches), and more complex than necessary.

The difficulty is that the decoders typically do not check all of the conditions (A), (B) and (C) of Theorem 1. The next two theorems explain why it is essential to make this check. Theorem 2 gives conditions on the polynomials σ_1 and ω that must always be satisfied, whether $\mathbf{R}$ is within RS distance r of a codeword, or not. Theorem 3, on the other hand, shows that the conditions imposed on σ_1 and ω in Theorem 2 are sufficient to guarantee the existence of a vector $\mathbf{R}$ which will produce these polynomials. Together, these two theorems show that there are $\mathbf{R}$s that will produce σ_1s and ωs satisfying some but not all of conditions (A), (B), and (C). Actually, condition (C) implies condition (B), but it is worthwhile to check condition (B) anyway since it is so easy to do so; if (B) fails, no further work is necessary. Theorems 2 and 3 show conditions (A) and (C) are independent, however, so that they both must be checked.

Theorem 2. For any word $\mathbf{R} \in \bar{F}^n$, the polynomials $\sigma_1(x)$ and $\omega(x)$ must satisfy the following three conditions:

(D) $$\deg(\sigma_1) \leq \mu$$

(E) $$\deg(\omega) \leq \nu$$

(F) $$\gcd(\sigma_1, \omega) = x^i$$

where x^i is the highest power of x dividing $\sigma_1(x)$.

Proof: Conditions (D) and (E) follow from the definition of the (a, b, μ, ν) problem. Condition (F) follows from Lemma C-3 in Appendix C. ■

Theorem 3. Conversely, given a set I_0 of t_0 erasure locations and any pair of polynomials $\sigma_1(x)$ and $\omega(x)$ satisfying conditions (D), (E), and (F), there exists a vector $\mathbf{R} \in \bar{F}^n$, and a choice of "erasure fills" Z_i (see Eq. 11) which will produce $\lambda\sigma_1(x)$ and $\lambda\omega(x)$ as the solution to the $(x^r, S_0(x), \mu, \nu)$ problem. Indeed, if $\sigma_1(x)$ and $\omega(x)$ are relatively prime, let $\mathbf{S} = [S_0, S_1, \ldots, S_{n-1}]$ be any vector such that

$$S_L + S_{L+1}x + \cdots + S_{L+r-1}x^{r-1} = \frac{\omega(x)}{\sigma(x)} \bmod x^r \tag{38}$$

where $\sigma(x) = \sigma_0(x)\,\sigma_1(x)$, and if the vector $\mathbf{R}' = [R'_0, \ldots, R'_{n-1}]$ is defined to be the inverse DFT of $\mathbf{S}$, i.e.,

$$R'_i = \frac{1}{n}\sum_{j=0}^{n-1} S_j \alpha^{-ij} \qquad \text{for } i = 0, \ldots, n-1 \tag{39}$$

and if $\mathbf{R}$ is defined by

$$R_i = \begin{cases} R'_i & \text{if } i \notin I_0 \\ * & \text{if } i \in I_0 \end{cases} \tag{40}$$

then if the RS decoding algorithm is applied to $\mathbf{R}$, (using the components of $\mathbf{R}'$ to fill in the erasures) $\sigma_1(x)$ and $\omega(x)$ will

be the error locator and error-and-erasure evaluator polynomials. Similarly, if $\gcd(\sigma_1, \omega) = x^j$ with $j > 0$, any **S** satisfying

$$S_L + S_{L+1}x + \cdots + S_{L+r-j-1}x^{r-j-1} = \frac{\omega(x)}{\sigma(x)} \bmod x^{r-j} \tag{41a}$$

$$S_L + S_{L+1}x + \cdots + S_{L+r-j-1}x^{r-j-1} + S_{L+r-j}x^{r-j} \neq \frac{\omega(x)}{\sigma(x)} \bmod x^{r-j+1} \tag{41b}$$

again with $\mathbf{R}'$ and **R** defined as in Eqs. (39) and (40), will do.

Proof: We distinguish two cases: $\gcd(\sigma_1, \omega) = 1$ and $\gcd(\sigma_1, \omega) \neq 1$. If $\gcd(\sigma_1, \omega) = 1$, and $S_0(x)$ is defined to be the polynomial $\omega(x)/\sigma_1(x) \bmod x^r$, then by Theorem C-4 in Appendix C and the example following it, (σ_1, ω) is the solution to the $(x^r, S_0(x), \mu, \nu)$ problem. Now if the decoding algorithm starts with **R**, and fills in the erasures to produce $\mathbf{R}'$, by Eq. (39) the syndrome polynomial $S(x)$ will be $S_L + S_{L+1}x + \cdots + S_{L+r-1}x^{r-1}$, which, by Eq. (38), is the same as $\omega(x)/\sigma(x) \bmod x^r$. Thus the modified syndrome $\sigma_0(x)\, S(x)$ will be $S_0(x) = \omega(x)/\sigma_1(x) \bmod x^r$, and, as we have seen, this means that the decoding algorithms will produce $s_1(x)$ and $\omega(x)$ as the error evaluator and error-and-erasure evaluator polynomials. The case when $\gcd(\sigma_1, \omega) \neq 1$ is handled similarly. ■

Corollary. If $t_0 \leqslant k$, then there is a vector $\mathbf{R}_0$ that produces (σ_1, ω) as the error locator polynomial and error-and-erasure evaluator with zeros as the erasure fills.

Proof: Let **R** be the vector defined by Eq. (40), and let **C** be any RS codeword that agrees with **R** on the set I_0. (There will be such a codeword, by the basic combinatorial property of RS codes, since $k > t_0$.) Then the vector $\mathbf{R} = \mathbf{R}_0 - \mathbf{C}$ will have the same syndrome as $\mathbf{R}_0$, and has zeros on the erasure set I_0. ■

References

[1] E. R. Berlekamp, *Algebraic Coding Theory*, Laguna Hills, California: Aegean Park Press, 1984. (Reprint of the 1968 McGraw-Hill original).

[2] R. E. Blahut, *Theory and Practice of Error Control Codes*, Reading, Massachusetts: Addison-Wesley, 1983.

[3] R. J. McEliece, *The Theory of Information and Coding*, Reading, Massachusetts: Addison-Wesley, 1977.

[4] R. J. McEliece and L. Swanson, "On the Decoder Error Probability for Reed-Solomon Codes," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 145–158, 1986.

[5] I. S. Reed, T. K. Troung, and R. L. Miller, "Simplified Algorithm for Correcting Both Errors and Erasures of Reed-Solomon Codes," *Proc. IEE*, vol. 126, No. 10, pp. 961–963, October 1979.

[6] D. Whiting, "Bit Serial Reed-Solomon Decoders in VLSI," Ph.D. thesis, California Institute of Technology, 1984.

[7] J. Yuen (ed.), *Deep Space Telecommunications Systems Engineering*, JPL Publication 82-76, Jet Propulsion Laboratory, Pasadena, California, 1982.

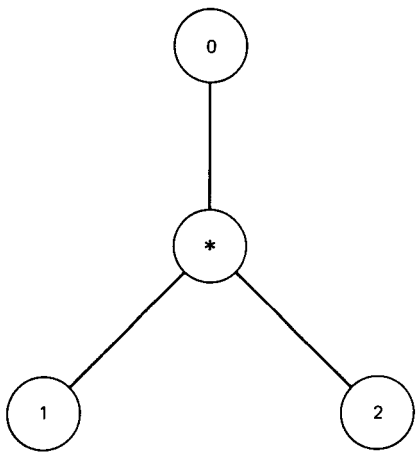


Fig. 1. Illustration of the RS decoding metric for GF(3).

Appendix A
The Discrete Fourier Transform

Let F be a field which contains a primitive nth root of unity α. (If the characteristic of F is finite, we assume that it does not divide n.) We first note

$$1 - x^n = \prod_{i=0}^{n-1} (1 - \alpha^i x) \tag{A-1}$$

This is because the polynomials on both sides of Eq. (A-1) have degree n, constant term 1, and roots α^{-i}, for $i = 0, 1, \ldots, n-1$.

Next, let

$$\mathbf{V} = (V_0, V_1, \ldots, V_{n-1}) \tag{A-2}$$

be an n-dimensional vector over F, and let

$$\widehat{\mathbf{V}} = (\widehat{V}_0, \widehat{V}_1, \ldots, \widehat{V}_{n-1}) \tag{A-3}$$

be its discrete Fourier transform (DFT), defined by

$$\widehat{V}_j = \sum_{i=0}^{n-1} V_i \alpha^{ij} \qquad \text{for } j = 0, 1, \ldots, n-1 \tag{A-4}$$

The components of $\mathbf{V}$ can be recovered from those of $\widehat{\mathbf{V}}$ via the inverse DFT

$$V_i = \frac{1}{n} \sum_{j=0}^{n-1} \widehat{V}_j \alpha^{-ij} \qquad \text{for } i = 0, 1, \ldots, n-1 \tag{A-5}$$

If we interpret the components of $\mathbf{V}$ and $\widehat{\mathbf{V}}$ as the coefficients of polynomials, i.e., if we define

$$V(x) = V_0 + V_1 x + \cdots + V_{n-1} x^{n-1} \tag{A-6}$$

and

$$\widehat{V}(x) = \widehat{V}_0 + \widehat{V}_1 x + \cdots + \widehat{V}_{n-1} x^{n-1} \tag{A-7}$$

then the DFT and IDFT relationships, Eqs. (A-4) and (A-5) become

$$\widehat{V}_j = V(\alpha^j) \tag{A-8}$$

and

$$V_i = \frac{1}{n} \widehat{V}(\alpha^{-i}) \tag{A-9}$$

Appendix B

Some Important Polynomials and the Fundamental Identity

Throughout this section I will denote a fixed subset of $\{0, 1, \ldots, n-1\}$. We associate several polynomials with this set. For example, the *locator polynomial* for I is

$$\sigma_I(x) = \prod_{i \in I} (1 - \alpha^i x) \tag{B-1}$$

The *co-locator polynomial* is

$$\tau_I(x) = \prod_{i \notin I} (1 - \alpha^i x) \tag{B-2}$$

In view of Eq. (A-1) we plainly have

$$1 - x^n = \sigma_I(x)\,\tau_I(x) \tag{B-3}$$

For each value of $i \in I$ we also define

$$\sigma_I^i(x) = \frac{\sigma_I(x)}{(1 - \alpha^i x)} = \prod_{\substack{j \in I \\ j \neq i}} (1 - \alpha^j x) \tag{B-4}$$

Finally, let $\mathbf{V} = (V_0, V_1, \ldots, V_{n-1})$ be a vector, such that I is a *support set* for $\mathbf{V}$, i.e., $V_i = 0$ if $i \notin I$. Then the $(\mathbf{V}, I)$ *evaluator polynomial* is defined as

$$\omega_{\mathbf{V},I}(x) = \sum_{i \in I} V_i \sigma_I^i(x) \tag{B-5}$$

We will need several lemmas about these polynomials.

Lemma B-1. For all $i, j \in I$,

$$\sigma_I^i(\alpha^{-j}) = \begin{cases} 0 & \text{if } j \neq i \\ \prod_{\substack{k \in I \\ k \neq i}} (1 - \alpha^{k-i}) & \text{if } j = i \end{cases}$$

In particular, $\sigma_I^I(\alpha^{-i}) \neq 0$.

Proof: Follows immediately from Eq. (B-4). ■

Lemma B-2. If $i \in I$, $\omega_{\mathbf{V},I}(\alpha^{-i}) = V_i \sigma_I^i(\alpha^{-i})$. In particular, $\omega_{\mathbf{V},I}(\alpha^{-i}) = 0$ if and only if $V_i = 0$.

Proof: This follows from Eq. (B-5) and Lemma B-1. ■

Lemma B-3. The polynomials $\sigma_I^i(x)$ are linearly independent, and therefore form a basis for the set of all polynomials of degree $< |I|$.

Proof: If

$$\sum_{i \in I} \lambda_i \sigma_I^i(x) = 0$$

on setting $x = \alpha^{-i}$, we would get by Lemma B-1, $\lambda_i \sigma_I^i(\alpha^{-i}) = 0$, but since (again by Lemma B-1) $\sigma_I^i(\alpha^{-i}) \neq 0$, this implies that $\lambda_i = 0$. The last statement of the lemma now follows from the facts that (a) each $\sigma_I^i(x)$ has degree exactly $|I| - 1$ and (b) there are exactly $|I|$ of them. ■

The next lemma deals with the *minimal* support set $I(\mathbf{V})$ of $\mathbf{V}$, which is defined by

$$I(\mathbf{V}) = \{i \in I : V_i \neq 0\} \tag{B-6}$$

In what follows, the corresponding polynomials will be denoted by $\sigma_{\mathbf{V}}(x)$, $\tau_{\mathbf{V}}(x)$, and $\omega_{\mathbf{V}}(x)$, rather than $\sigma_{I(\mathbf{V})}(x)$, etc.

Lemma B-4.

$$\gcd(\sigma_I(x), \omega_{\mathbf{V},I}(x)) = \prod_{I - I(\mathbf{V})} (1 - \alpha^i x)$$

In particular, $\gcd(\sigma_{\mathbf{V}}(x), \omega_{\mathbf{V}}(x)) = 1$.

Proof: If $\omega_{\mathbf{V},I}(x)$ had a factor in common with $\sigma_I(x)$, then by Eq. (B-3) $\omega_{\mathbf{V},I}(\alpha^{-i}) = 0$ for some $i \in I$. But by Lemma B-2, this is true if and only if $V_i = 0$, i.e., if $i \in I - I(\mathbf{V})$. ■

We note that for any $\mathbf{V}$ and support set I, $\sigma_I(x)$ divides $1 - x^n$ and $\deg \omega_{\mathbf{V},I}(x) < \deg \sigma_I(x)$. The following theorem is a kind of a converse to this.

Theorem B-5. Suppose $\sigma(x)$ and $\omega(x)$ are polynomials such that $\sigma(x) | 1 - x^n$ and $\deg(\omega) < \deg(\sigma)$. Then there exists a vector $\mathbf{V}$ and a support set I for $\mathbf{V}$ such that

$$\sigma(x) = \lambda \sigma_I(x) \tag{B-7}$$

$$\omega(x) = \lambda\omega_{\mathbf{V},I}(x) \tag{B-8}$$

for a nonzero constant λ. Furthermore, if in addition $\gcd(\sigma(x), \omega(x)) = 1$, then in fact there exists a vector $\mathbf{V}$ such that

$$\sigma(x) = \lambda\sigma_{\mathbf{V}}(x) \tag{B-9}$$

$$\omega(x) = \lambda\omega_{\mathbf{V}}(x) \tag{B-10}$$

Proof: Suppose $\sigma(x) \mid 1 - x^n$. Then Eq. (B-7) must hold for some subset I of $\{0, 1, \ldots, n-1\}$ and some nonzero constant λ. Since $\deg(\omega) < \deg(\sigma) = \deg(\sigma_I)$, and since the polynomials $\sigma_I^i(x)$ are linearly independent by Lemma B-3,

$$\omega(x) = \lambda \sum_{i\in I} u_i \sigma_I^i(x) \tag{B-11}$$

for certain constants u_i. Thus if we define

$$V_i = \begin{cases} u_i & \text{if } i \in I \\ 0 & \text{if } i \notin I \end{cases} \tag{B-12}$$

Eq. (B-10) follows on comparing Eq. (B-11) to Eq. (B-5). Finally, by Lemma B-4, $\gcd(\sigma_I(x), \omega_{\mathbf{V},I}(x)) = 1$ if and only if $I = I(\mathbf{V})$, and so if $\gcd(\sigma(x), \omega(x)) = 1$, Eqs. (B-7) and (B-8) become Eqs. (B-9) and (B-10). ■

The next theorem is the most important result in this section.

Theorem B-6. If I is a support set for $\mathbf{V}$, then the polynomials $V(x)$, $\sigma_I(x)$, and $\omega_{\mathbf{V},I}(x)$ satisfy

$$\sigma_I(x)\,\widehat{V}(x) = \omega_{\mathbf{V},I}(x)\,(1 - x^n) \tag{B-13}$$

Proof: Using the definitions in Eqs. (A-4) and (A-7), together with the fact that I is a support set for $\mathbf{V}$, we find that

$$\widehat{\mathbf{V}}(x) = \sum_{i\in I} V_i \sum_{j=0}^{n-1} x^j \alpha^{ij} \tag{B-14}$$

According to Eq. (B-4), $\sigma_I(x) = \sigma_I^i(x)\,(1 - \alpha^i x)$ for all $i \in I$, and so from Eq. (B-14) we have

$$\sigma_I(x)\,\widehat{\mathbf{V}}(x) = \sum_{i\in I} V_i \sigma_I^i(x)\,(1 - \alpha^i x) \sum_{j=0}^{n-1} x^j \alpha^{ij}$$

$$= \sum_{i\in I} V_i \sigma_I^i(x)\,(1 - x^n)$$

$$= \omega_{\mathbf{V},I}(x)\,(1 - x^n)$$ ■

The following Corollary to Theorem B-6 tells us how to reconstruct the nonzero components of $\mathbf{V}$ from $\sigma_I(x)$ and $\omega_{\mathbf{V},I}(x)$. It involves the *formal derivative* $\sigma_I'(x)$ of the polynomial $\sigma_I(x)$.

Corollary B-7. If I is a support set for $\mathbf{V}$, then for each $i \in I$, we have

$$V_i = -\alpha^i \frac{\omega_{\mathbf{V},I}(\alpha^{-i})}{\sigma_I'(\alpha^{-i})} \tag{B-15}$$

Proof: If we differentiate the fundamental identity in Eq. (B-13) we obtain

$$\sigma_I(x)\,\widehat{V}'(x) + \sigma_I'(x)\,\widehat{V}(x) = \omega_{\mathbf{V},I}(x)\,(-nx^{n-1}) + \omega_{\mathbf{V},I}'(x)\,(1 - x^n) \tag{B-16}$$

Note that if $x = \alpha^{-i}$ with $i \in I$, from Eqs. (B-3) and (A-1) we see that both $\sigma_I(x)$ and $1 - x^n$ vanish. Thus if $x = \alpha^{-i}$, Eq. (B-16) becomes

$$\sigma_I'(\alpha^{-i})\,\widehat{V}(\alpha^{-i}) = -n\alpha^i \omega_{\mathbf{V},I}(\alpha^{-i}) \tag{B-17}$$

But from Eq. (A-9), $\widehat{V}(\alpha^{-i}) = nV_i$. This fact, combined with Eq. (B-17), completes the proof. ■

Corollary B-8. $\gcd(\widehat{V}(x), 1 - x^n) = \tau_{\mathbf{V}}(x)$.

Proof: From Eq. (B-3) with $I = I(\mathbf{V})$, we have $1 - x^n = \sigma_{\mathbf{V}}(x)\,\tau_{\mathbf{V}}(x)$. Then, if we divide both sides of Eq. (B-13) by $\sigma_{\mathbf{V}}(x)$, we get $\widehat{V}(x) = \omega_{\mathbf{V}}(x)\,\tau_{\mathbf{V}}(x)$. Since by Lemma B-4, $\gcd(\sigma_{\mathbf{V}}(x), \omega_{\mathbf{V}}(x)) = 1$, the Corollary follows. ■

Now we can prove a kind of converse to Theorem B-6.

Theorem B-9. Suppose that the vector $\mathbf{V}$ is given, and that for certain polynomials $\sigma(x)$ and $\omega(x)$ we have

$$\sigma(x)\,\widehat{\mathbf{V}}(x) = \omega(x)\,(1 - x^n) \tag{B-18}$$

Then there exists a polynomial $\lambda(x)$ such that

$$\sigma(x) = \lambda(x)\,\sigma_V(x) \tag{B-19}$$

$$\omega(x) = \lambda(x)\,\omega_V(x) \tag{B-20}$$

Proof: By Eq. (B-18) we have

$$\sigma(x)\,\hat{V}(x) \equiv 0 \pmod{1 - x^n} \tag{B-21}$$

This implies that

$$\sigma(x) \equiv 0 \left(\mod \frac{1 - x^n}{\gcd(1 - x^n, \hat{V}(x))}\right) \tag{B-22}$$

But by Corollary B-8, $\gcd(1 - x^n, \hat{V}(x)) = \tau_V(x)$, and from Eq. (B-3),

$$\frac{(1 - x^n)}{\tau_V(x)} = \sigma_V(x)$$

and so Eq. (B-22) is equivalent to Eq. (B-19), for a suitable polynomial $\lambda(x)$. Then Eq. (B-18) becomes

$$\lambda(x)\,\sigma_V(x)\,\hat{V}(x) = \omega(x)\,(1 - x^n) \tag{B-23}$$

but multiplying Eq. (B-13) by $\lambda(x)$ we obtain

$$\lambda(x)\,\sigma_V(x)\,\hat{V}(x) = \lambda(x)\,\omega_V(x)\,(1 - x^n) \tag{B-24}$$

Comparing Eq. (B-23) to Eq. (B-24), we see that

$$\omega(x) = \lambda(x)\,\omega_V(x) \tag{B-25}$$

as asserted. This completes the proof of Theorem B-9. ■

The next results in this section deal with homogeneous difference equations (HDEs). We say that the infinite sequence $u_0, u_1, \ldots$ satisfies a dth-order HDE if there exist constants $\sigma_0, \ldots, \sigma_d$, with $\sigma_0 \neq 0$ and $\sigma_d \neq 0$ such that

$$\sum_{k=0}^{d} \sigma_k u_{j-k} = 0 \qquad \text{for } j \geq d \tag{B-26}$$

The polynomial $\sigma(x) = \sigma_0 + \cdots + \sigma_d x^d$ is called the *characteristic polynomial* of the HDE, and the degree d of $\sigma(x)$ is called its *order*. If we define

$$\omega_j = \sum_{k=0}^{j} \sigma_k u_{j-k} \qquad \text{for } j = 0, \ldots, d-1 \tag{B-27}$$

and $\omega(x) = \omega_0 + \cdots + \omega_{d-1} x^{d-1}$, then it follows from Eqs. (B-26) and (B-27) that (u_j) satisfies an HDE with characteristic polynomial $\sigma(x)$ if and only if

$$\sigma(x)\,u(x) = \omega(x) \tag{B-28}$$

where $\deg(\omega(x)) < \deg(\sigma(x))$. In particular, the sequence (u_j) is periodic of period n, i.e., $u_j = u_{j-n}$ for $j \geq n$, if and only if there is a polynomial $\Omega(x)$ of degree $< n$ such that

$$(1 - x^n)\,u(x) = \Omega(x) \tag{B-29}$$

where $\deg \Omega < n$. The following theorem is needed in the discussion of the frequency-domain decoder. It assumes that (u_j) is a sequence that satisfies a dth-order HDE with characteristic polynomial $\sigma(x)$, as described by Eq. (B-28).

Theorem B-10. If $\sigma(x)$ divides $1 - x^n$, then the sequence (u_j) has period n. Conversely, if (u_j) has period n and if $\sigma(x) = \sigma_0(x)\,\sigma_1(x)$, where $\sigma_0(x)$ divides $1 - x^n$ and $\gcd(\sigma_1(x), \omega(x)) = 1$, then $\sigma(x)$ divides $1 - x^n$. In particular, if $\gcd(\sigma(x), 1 - x^n) = 1$, then $\sigma(x) \mid 1 - x^n$.

Proof: If $\sigma(x)$ divides $1 - x^n$, then $1 - x^n = \sigma(x)\,\tau(x)$ for some polynomial $\tau(x)$. If we multiply both sides of Eq. (B-28) by $\tau(x)$, we obtain $(1 - x^n)\,u(x) = \omega(x)\,\tau(x)$. But $\deg(\omega(x) \cdot \tau(x)) < \deg(\sigma(x)\,\tau(x)) = n$, and so by Eq. (B-29) (u_j) has period n.

Conversely, if Eq. (B-29) holds, and we multiply Eq. (B-28) by $1 - x^n$ and Eq. (B-29) by $\sigma(x)$, then we find that $\Omega(x) \cdot \sigma(x) = \omega(x)\,(1 - x^n)$. Therefore $\sigma(x) \mid \omega(x)\,(1 - x^n)$. Since $\sigma(x) = \sigma_0(x)\,\sigma_1(x)$ and $\sigma_0(x) \mid 1 - x^n$, it follows that $\sigma_1(x) \mid \omega(x)\,(1 - x^n)/\sigma_0(x)$. But $\gcd(\sigma_1(x), \omega(x)) = 1$, and this means that $\sigma_1(x) \mid (1 - x^n)/\sigma_0(x)$, which implies that $\sigma(x) = \sigma_0(x)\,\sigma_1(x) \mid 1 - x^n$. ■

Having briefly discussed homogeneous difference equations, we are now in a position to discuss *circular homogeneous difference equations* (CHDEs). We say that the finite sequence $(u_0, \ldots, u_{n-1})$ satisfies a dth-order CHDE if there are constants $\sigma_0, \sigma_1, \ldots, \sigma_d$ with $\sigma_0 \neq 0$ and $\sigma_d \neq 0$, such that

$$\sum_{k=0}^{d} \sigma_k u_{j-k} = 0 \qquad \text{for } j = 0, \ldots, n-1 \tag{B-30}$$

where the subscripts must be interpreted mod n. The polynomial $\sigma(x) = \sigma_0 + \sigma_1 x + \cdots + \sigma_d x^d$ is called the characteristic polynomial of the CHDE, and d is its order. If we define

$u(x) = u_0 + u_1 x + \cdots + u_{n-1}x^n$, then Eq. (B-30) holds if and only if

$$\sigma(x)\,u(x) \equiv 0 \pmod{1 - x^n} \tag{B-31}$$

Equivalently, $(u_0, \ldots, u_{n-1})$ satisfies a CHDE if and only if there is a polynomial $\omega(x)$ such that

$$\sigma(x)\,u(x) = \omega(x)\,(1 - x^n) \tag{B-32}$$

Plainly, the $\sigma(x)$ of smallest degree such that Eq. (B-32) holds is

$$\sigma_{\min}(x) = \frac{1 - x^n}{\gcd(u(x), 1 - x^n)} \tag{B-33}$$

which is a divisor of $1 - x^n$. Thus Theorem B-6 says that $\hat{\mathbf{V}}$ satisfies a CHDE of order $|I|$, where I is any support set for $\mathbf{V}$. Conversely, Theorem B-8 says that $\hat{\mathbf{V}}$ does not satisfy a CHDE of order lower than $|I(\mathbf{V})|$. But we know from Eq. (B-6) that $|I(\mathbf{V})|$ = weight $(\mathbf{V})$ and so we have proved Theorem B-11.

Theorem B-11. The weight of $\mathbf{V}$ is the degree of the least-order CHDE satisfied by $\hat{\mathbf{V}}$.

Appendix C
The $(a(x), b(x), \mu, \nu)$ Problem

Given polynomials $a(x)$, $b(x)$, with $\deg(b) < \deg(a) = m$, and nonnegative integers μ, ν with $\mu + \nu = m - 1$, consider the set $S = S(a, b, \mu, \nu)$ of all pairs of polynomials $(\sigma(x), \omega(x))$ such that

$$\deg(\sigma) \leqslant \mu \qquad \deg(\omega) \leqslant \nu \tag{C-1}$$

$$\sigma(x)\, b(x) \equiv \omega(x) \pmod{a(x)} \tag{C-2}$$

Theorem C-1. If $\mu + \nu = m - 1$, the set $S(a, b, \mu, \nu)$ is not empty. Indeed, there exists a pair $(\sigma_0, \omega_0) \in S(a, b, \mu, \nu)$ such that every pair $(\sigma(x), \omega(x)) \in S(a, b, \mu, \nu)$ is of the form

$$\sigma(x) = k(x)\, \sigma_0(x) \tag{C-3}$$

$$\omega(x) = k(x)\, \omega_0(x) \tag{C-4}$$

Furthermore, (σ_0, ω_0) is unique up to multiplication by scalars. We summarize this by saying that (σ_0, ω_0) "solves the $(a(x), b(x), \mu, \nu)$ problem."

Proof: A proof is given in [3], Theorem 8.5, where it is also pointed out that Euclid's algorithm can be used to find (σ_0, ω_0). Specifically, if one applies Euclid's algorithm as described there to the pair $(a(x), b(x))$, and stops when the degree of the remainder $r_j(x)$ becomes $\leqslant \nu$ for the first time, then $(t_j(x), r_j(x))$ is the solution to the $(a(x), b(x), \mu, \nu)$ problem. ■

Lemma C-2. $(\sigma, \omega) \in S(a, b, \mu, \nu)$ solves the (a, b, μ, ν) problem if and only if there exists a polynomial $\tau(x)$ such that

$$\omega = \sigma b + \tau a \tag{C-5}$$

where

$$\gcd(\omega, \sigma, \tau) = 1 \tag{C-6}$$

Proof: Suppose that (σ, ω) solves the problem. Then by Eq. (C-2), there exists a polynomial $\tau(x)$ such that Eq. (C-5) holds. If $\omega(x)$, $\sigma(x)$, and $\tau(x)$ had a common factor $d(x)$, then with $\omega' = \omega/d$, $\sigma' = \sigma/d$, and $\tau' = \tau/d$, Eq. (C-5) implies $\omega' = \sigma' b + \tau' a$, which means $\sigma' b \equiv \omega' \pmod{a}$ is a smaller degree solution to Eqs. (C-1) and (C-2), contradicting the minimality of (σ, τ).

Conversely, suppose Eqs. (C-5) and (C-6) hold, and that (σ_0, ω_0) solves the (a, b, μ, ν) problem. Then by Theorem C-1, $\omega = k\omega_0$, $\sigma = k\sigma_0$, and Eq. (C-5) becomes

$$k\omega_0 = k\sigma_0 b + \tau a \tag{C-7}$$

But since $(\sigma_0, \omega_0) \in S(a, b, \mu, \nu)$, we know that $\omega_0 = \sigma_0 b + \tau_0 a$ for some polynomial $\tau_0(x)$, and so we have

$$k\omega_0 = k\sigma_0 b + k\tau_0 a \tag{C-8}$$

Comparing Eqs. (C-7) and (C-8), we see that $\tau = k\tau_0$, and so $k \mid \gcd(\omega_1, \sigma_1, \tau_1)$. This implies by Eq. (C-6) that k is a scalar and so (σ, ω) is a scalar multiple of (σ_0, ω_0), i.e., (σ_1, ω_1) solves the (a, b, μ, ν) problem. ■

Lemma C-3. If (σ_0, ω_0) solves the (a, b, μ, ν) problem, then

$$\gcd(\sigma_0, \omega_0) = \gcd(\sigma_0, a) \tag{C-9}$$

Proof: By Lemma C-2 we know that

$$\omega_0 = \sigma_0 b + \tau_0 a \tag{C-10}$$

with

$$\gcd(\omega_0, \sigma_0, \tau_0) = 1 \tag{C-11}$$

Now by Eq. (C-10) any common divisor of σ_0 and a must divide ω_0, i.e., $\gcd(\sigma_0, a) \mid \gcd(\sigma_0, \omega_0)$. On the other hand, Eq. (C-10) also says that any common divisor of σ_0 and ω_0 must divide $\tau_0 a$ and so by Eq. (C-11) must divide a. Thus $\gcd(\sigma_0, \omega_0) \mid \gcd(\sigma_0, a)$. ■

Theorem C-4. Conversely, given polynomials $a(x)$, $\sigma_0(x)$, and $\omega_0(x)$ such that Eqs. (C-1) and (C-9) hold, there exists a polynomial $b(x)$ of degree $\leqslant m - 1$ such that (σ_0, ω_0) solves the $(a(x), b(x), \mu, \nu)$ problem.

Proof: Let $d(x) = \gcd(\sigma_0(x), \omega_0(x)) = \gcd(\sigma_0(x), a(x))$, and $\sigma_1 = \sigma_0/d$, $\omega_1 = \omega_0/d$, $a_1 = a/d$. Since $\gcd(\sigma_1, a_1) = 1$ we can define

$$b' = \frac{\omega_1}{\sigma_1} \bmod a_1 \tag{C-12}$$

It follows that

$$\sigma_1 b' \equiv \omega_1 \pmod{a_1} \tag{C-13}$$

i.e.,

$$\omega_1 = \sigma_1 b' + \tau_1 a_1 \tag{C-14}$$

for a suitable polynomial $\tau_1(x)$. We note that

$$\gcd(\sigma_1, \tau_1) = 1 \tag{C-15}$$

since a common factor of σ_1 and τ_1 would by Eq. (C-14) also divide ω_1; but $\gcd(\sigma_1, \omega_1) = 1$. We now distinguish two cases, according to whether d and τ_1 have a factor in common or not.

Case 1: $\gcd(d, \tau_1) = 1$. In this case if we multiply Eq. (C-14) by d we obtain

$$\omega_0 = \sigma_0 b' + \tau_1 a \tag{C-16}$$

with $\gcd(\sigma_0, \omega_0, \tau_1) = \gcd(d, \tau_1) = 1$, and so by Lemma C-2, (σ_0, ω_0) solves the (a, b', μ, ν) problem, where b' is defined by Eq. (C-12).

Case 2: $\gcd(d, \tau_1) \neq 1$. In this case, let $\lambda(x)$ be the product of all the irreducible polynomials which divide p but do not divide τ_1, i.e.,

$$\lambda = \prod \{p : p \text{ is irreducible}, p \mid d, p \nmid \tau_1\} \tag{C-17}$$

Next, we define

$$\begin{aligned} b &= b' + \lambda a_1 \\ \tau_0 &= \tau_1 - \lambda \sigma_1 \end{aligned} \tag{C-18}$$

Then from Eq. (C-14) it follows that

$$\omega_1 = \sigma_1 b + \tau_0 a_1 \tag{C-19}$$

If p is an irreducible divisor of d, then p cannot divide τ_0, because if $p \mid \tau_1$ then $p \nmid \sigma_1$ by Eq. (C-15) and $p \nmid \lambda$ by Eq. (C-17), so that $p \nmid \tau_0 = \tau_1 - \lambda\sigma_1$. On the other hand, if $p \nmid \tau_1$, then by Eq. (C-17) $p \mid \lambda$ and so again $p \nmid \tau_0 = \tau_1 - \lambda\sigma_1$. This shows that $\gcd(d, \tau_0) = 1$. Thus if we multiply Eq. (C-19) by d we obtain

$$\omega_0 = \sigma_0 b + \tau_0 a \tag{C-20}$$

with $\gcd(\omega_0, \sigma_0, \tau_0) = \gcd(d, \tau_0) = 1$, and so by Lemma C-2, (σ_0, ω_0) solves the (a, b, μ, ν) problem, with b defined in Eq. (C-18). ■

Example: Suppose $a(x) = x^m$. Then the construction of Theorem C-4 simplifies considerably. Indeed, suppose we are given polynomials $\sigma_0(x)$ and $\omega_0(x)$ such that Eqs. (C-1) and (C-9) hold, with $a(x) = x^m$. Then $\gcd(\sigma_0, \omega_0) = \gcd(\sigma_0, x^m) = x^j$ for some value of j. If $b(x)$ is such that Eq. (C-2) holds, then dividing by x^j we obtain

$$\sigma_1 b \equiv \omega_1 \pmod{x^{m-j}} \tag{C-21}$$

where $\gcd(\sigma_1, x^{m-j}) = 1$. Thus if (σ_0, ω_0) is to solve the (x^m, b, μ, ν) problem, then it must be true that

$$b(x) \equiv \frac{\omega_1}{\sigma_1} \pmod{x^{m-j}} \tag{C-22}$$

If $j = 0$, i.e., if $\gcd(\sigma_0, \omega_0) = 1$, then Eq. (C-22) is both necessary and sufficient. On the other hand, if $j \geq 1$, it is easy to see that Lemma C-2 implies that (σ_0, ω_0) solves the (x^m, b, μ, ν) problem if and only if Eq. (C-22) holds and if in addition

$$b(x) \not\equiv \frac{\omega_1}{\sigma_1} \pmod{x^{m-j+1}} \tag{C-23}$$

Thus if we expand ω_1/σ_1 as a power series, viz.

$$\frac{\omega_1}{\sigma_1} = a_0 + a_1 x + \cdots \tag{C-24}$$

then any $b(x) = b_0 + b_1 x + \cdots + b_{m-1} x^{m-1}$ will do, provided that

$$\begin{aligned} b_k &= a_k \qquad \text{for } k = 0, \ldots, m-j-1 \\ b_{m-j} &\neq a_{m-j} \end{aligned} \tag{C-25}$$

Performance of Efficient Q-Switched Diode-Laser-Pumped Nd:YAG and Ho:YLF Lasers for Space Applications

W. K. Marshall, K. Cowles, and H. Hemmati
Communications Systems Research Section

Solid-state lasers pumped by continuous-wave diode lasers can be Q-switched to obtain high-peak-power output pulses. In this article, the dependence of laser-pulse energy, average output power, peak power, and pulse width on pulse-repetition frequency in Q-switched Nd:YAG and Ho:YLF lasers is determined and compared. At low pulse-repetition rates, the much longer upper-state lifetime in Ho:YLF gives a distinct advantage. At higher pulse rates, the overall laser efficiency and the stimulated emission cross section are more important parameters, leading to an advantage for Nd:YAG. The results are of significance for designing lasers for use in space optical communications and remote sensing systems.

I. Introduction

Diode-laser-pumped solid-state lasers such as neodymium: yttrium aluminum garnet (Nd:YAG) and holmium: yttrium lithium flouride (Ho:YLF) are prime candidates for laser-light sources for use in deep-space optical communications and other applications. Due to the long upper-state lifetimes of the Nd^{3+} and Ho^{3+} ions doped into the YAG and YLF crystals, the energy storage capacity of such laser materials is quite high. Using a technique such as Q switching or cavity dumping, the stored energy can be extracted in the form of short laser pulses with high peak power. (This form is optimal for direct-detection optical communications use.) The pulse-repetition rate is controllable using an electro-optical or acoustical-optical device for Q switching.

A laser Q switch is effectively an intracavity shutter—when the Q switch is closed, optical pumping continues, but stimulated emission does not occur. During this time, energy is continuously stored in the upper laser level. When the Q switch is suddenly opened, the stored energy is released in the form of a "giant pulse" of laser light, depleting the upper laser level. The next laser pulse occurs only after a sufficient population of ions is again placed in the upper laser level.

Efficient diode-pumped continuous-wave (CW) Nd:YAG [1] and Ho:YLF [2] lasers were first demonstrated at JPL. In this article, the expected output power and pulse characteristics of diode-pumped, Q-switched (pulsed) Nd:YAG and Ho:YLF lasers are analyzed. In Section II below, the basic theory needed to calculate the results of interest is given. In Section III, the factors relevant to a comparison of Q-switched lasers are explicitly considered. In Section IV, the results are applied to specific laser examples.

II. Basic Theory

The most important factor determining the pulse shape and pulse energy in a Q-switched laser is the population inversion

density[1] of the lasing medium just prior to the opening of the Q switch. In this section, we summarize the basic theory [3] for determining population inversion, power output, and pulse width in a CW-pumped, Q-switched laser.

In a series of periodic laser pulses, the population inversion density of the initial state (i.e., just before a laser pulse), n_i, depends on the inversion density of the final state (i.e., just after the preceding laser pulse), n_f, and on the amount of pumping that occurs during the period between the laser pulses, according to the equation

$$n_i = n_\infty - (n_\infty - n_f)e^{-1/\tau_s f} \tag{1}$$

where f is the pulse-repetition frequency and τ_s is the upper-laser-level spontaneous decay time. This equation is for continuous pumping at a uniform rate. The asymptotic density, n_∞, depends on the pumping rate and is the maximum achievable inversion density (approached when $1/f \gg \tau_s$). The value of n_∞ can be calculated from knowledge of the CW output power, P_{cw}, obtained when the cavity Q is maintained at its maximum value:

$$n_\infty = \frac{P_{cw}\tau_s}{\eta h\nu V} + n_{th} \tag{2}$$

Here, $h\nu$ is the photon energy, and V is the effective lasing volume, n_{th} is the threshold inversion density, and η is the output coupling factor. The latter two quantities are given [4], [5] by

$$n_{th} = \frac{1}{\sigma}\left[\frac{1}{l}\ln\left(\frac{1}{\xi\sqrt{r_1 r_2}}\right) + \beta\right] \tag{3}$$

$$\eta = \frac{\ln r_1}{\ln \xi^2 + \ln r_1 + \ln r_2 - 2\beta l} \tag{4}$$

for stimulated emission cross section σ, length of laser rod l, Q-switch maximum single pass transmission ξ, reflectivities of output and rear mirrors, r_1 and r_2, respectively, and laser-rod-loss coefficient β.

[1]The inversion density, n, is the difference between the population density of the upper laser level and that of the lower laser level.

The population inversion just after the laser pulse, n_f, is given[2] by the (transcendental) equation

$$n_i - n_f = n_{th}\ln(n_i/n_f) \tag{5}$$

This equation, together with Eq. (1) above, determines the values for n_i and n_f for a given set of conditions. Once the change in the population inversion $n_i - n_f$ is known, the energy per pulse is given simply by

$$E_{pulse} = (n_i - n_f)\eta h\nu V \tag{6}$$

The average power output from the laser is then

$$P_{avg} = E_{pulse} f \tag{7}$$

The pulse shape is determined by the laser rate equations. From those equations, the peak laser power is found [3] to be

$$P_{peak} = \frac{Vh\nu \ln\left(\frac{1}{r_1}\right)}{\frac{1}{2}t_R}\left\{n_i - n_{th}\left[1 + \ln\left(\frac{n_i}{n_{th}}\right)\right]\right\} \tag{8}$$

where t_R is the round trip cavity time. Since $t_R = 2L/c$ where L is the cavity optical path length and c is the speed of light, a significant consequence of the equation for P_{peak} is that the peak output power of a Q-switched laser is inversely proportional to the cavity length.

III. Comparison of Lasers

Consider first the energy per pulse, E_{pulse}. Use of Eqs. (1), (2), and (6) gives

$$E_{pulse} = P_{cw}\tau_s\left(1 - e^{-1/\tau_s f}\right)\left(1 + \frac{\hat{n}_f}{\hat{n}_\infty}\right) \tag{9}$$

where $\hat{n}_\infty = (n_\infty - n_{th})/n_{th}$ and $\hat{n}_f = (n_{th} - n_f)/n_{th}$. The variables $\hat{n}_\infty$ and $\hat{n}_f$ are normalized versions of their "unhatted" counterparts, and have ranges $0 < \hat{n}_\infty < \infty$, and $0 < \hat{n}_f < 1$ (i.e., the final inversion density is less than the CW laser threshold).

The exact value of $\hat{n}_f$ can be determined only by solving Eqs. (1) and (5) numerically. Alternatively, we note from the

[2]Subject to the approximation that the effect of pumping *during* the laser pulse is negligible.

form[3] of Eq. (5) that $\hat{n}_f < \hat{n}_i$. Since $\hat{n}_i < \hat{n}_\infty$ by definition, $0 < \hat{n}_f < \hat{n}_\infty$ and hence the rightmost factor in Eq. (9) ranges at most between 1 and 2.

Thus, within a factor close to unity, Eq. (9) says that the energy per pulse for a Q-switched laser depends only on the laser's CW output power level, the upper-state lifetime for the laser, and on the pulse-repetition frequency, and not on other factors such as the laser-stimulated emission cross section or the passive characteristics of the laser cavity.

For high pulse rates $f \gg 1/\tau_s$, Eq. (9) reduces to $E_{pulse} \simeq P_{cw}/f$, i.e., effectively the laser's CW power is collected over the pump time $1/f$ and emitted as a short pulse. For low pulse rates, the pulse energy saturates as the pumping time becomes long compared to the "storage time" τ_s. In this latter case, Eq. (9) reduces to $E_{pulse} \simeq P_{cw}\tau_s$.

Also of significance in comparing two lasers are the respective pulse widths. The laser pulse width limits the modulation alphabet size, limits the ability to reduce background light by narrowing the signal slot width, and (in the extreme) limits the maximum achievable Q-switched pulse frequency. The pulse shape can be determined exactly only by numerically integrating the laser-rate equations. Here we calculate only an estimate (actually a lower bound) for the laser pulse width given by

$$t_{pulse} \cong \frac{E_{pulse}}{P_{peak}} \tag{10}$$

where E_{pulse} and P_{peak} are given by Eqs. (6) and (8) above. For high pulse rates such that $f \gg 1/\tau$ and $n_i - n_{th} \ll n_{th}$, this reduces to

$$t_{pulse} \cong \frac{t_R\, \eta h\nu V f}{2\sigma l P_{cw}} \tag{11}$$

i.e., the pulse width is directly proportional to the pulse rate. For low pulse rates, $f \ll 1/\tau_s$, the pulse width approaches a constant (minimum), as both the pulse energy and the peak power saturate. This constant value is given by

$$t_{pulse}^{(min)} = \frac{t_R \eta}{2\ln(1/r_1)} \frac{\hat{n}_\infty}{\hat{n}_\infty - \ln(\hat{n}_\infty + 1)} \tag{12}$$

where $\hat{n}_\infty$ is defined above.

[3]Equation (5) can be rewritten as $(1 + \hat{n}_i) - \ln(1 + \hat{n}_i) = (1 - \hat{n}_f) - \ln(1 - \hat{n}_f)$, where $\hat{n}_i = (n_i - n_{th})/n_{th}$.

IV. Specific Laser Examples

Now consider and compare Nd:YAG and Ho:YLF lasers characterized by a single-mode CW output power of 100 mW and a cavity length of 3.5 cm. Other laser parameters are given in Table 1. The most significant difference between the two lasers is that the lifetime of the upper laser level, τ_s, is about 50 times longer in Ho:YLF than in Nd:YAG. The parameters for the 100-mW lasers were chosen to represent the (CW) lasers reported in [1] and [2]. Scaling to higher powers is considered briefly at the end of this section.

Figure 1 shows a plot of the energy per pulse, E_{pulse}, (given by Eq. (9) with the factor $(1 + \hat{n}_f/\hat{n}_\infty)$ set equal to unity) as a function of the pulse-repetition frequency. For both lasers, the initial state inversion density saturates as the pumping time $(1/f)$ begins to be long compared to the respective upper-state lifetimes. The saturation value of the pulse energy is $(\tau_s \times 100\ \text{mW})$; hence, for low pulse rates, the Ho:YLF laser pulses are about 50 times larger than those for Nd:YAG. For high pulse rates, the pulse energies become asymptotically equal. (At intermediate pulse rates, the Ho:YLF pulse energy is always larger.)

Figure 2 shows the laser average output power, P_{avg}, versus frequency. The average power goes to zero for low repetition rates, and approaches the CW laser power (100 mW) at high pulse-repetition rates. Again, values are always higher for the Ho:YLF laser.

Figure 3 shows the estimate on the laser pulse width given by Eq. (10). At low pulse-repetition rates, the Ho:YLF pulse width is about three times smaller than the Nd:YAG pulse width, due mainly to the effect of a higher $\hat{n}_\infty$ for the Ho:YLF. For higher pulse rates (above $\sim 10^3$ per second), the pulse widths are controlled by Eq. (11); here, since $h\nu$ and $1/\sigma$ are smaller for Nd:YAG than for Ho:YLF, the Nd:YAG pulse width is smaller by a factor of about 4. In the frequency range of 10^5 to 10^6 per second (upper right corner of Fig. 3), the calculated lower bound on the pulse width for both lasers exceeds the pulse interval $1/f$—the laser will not operate in a simple Q-switched pulse mode at those frequencies.

Now consider a simple scaling example—two 1-W lasers. (Laser parameters are the same as those given in Table 1, except $P_{cw} = 1$ W.) It can be seen from Eq. (9) that E_{pulse} scales linearly with P_{cw} (subject to the validity of the approximation $\hat{n}_f \ll \hat{n}_\infty$). Therefore, values of E_{pulse} (and P_{avg}) for 1-W lasers are a factor of 10 larger than those for the 100-mW lasers shown in Figs. 2 and 3. Values for t_{pulse} are smaller for the 1-W lasers than those shown in Fig. 3 for 100-mW lasers. For high pulse rates, Eq. (11) gives a reduction in pulse width by a factor of 10. At lower pulse rates, the reduction is not a linear

factor–calculated values of $t_{pulse}^{(min)}$ for the 1-W lasers are 1.6 ns and 3.4 ns for Ho:YLF and Nd:YAG, respectively.

V. Conclusions

The Ho:YLF laser shows promise for use in low-data-rate optical communications systems, such as would likely be used in an Earth-to-spacecraft uplink in an all-optical communication system. At pulse repetition rates below 10 kHz, Ho:YLF offers higher performance than Nd:YAG in terms of higher energy per pulse for similar CW lasers. The upper-state lifetime, τ_s, is the dominant parameter in determining Q-switched pulse energy in this regime. Below 1 kHz, the Ho:YLF also offers shorter laser pulse widths. The large value of τ_s makes it easy to pump Ho:YLF far above threshold, resulting in short pulse widths. At higher pulse rates, τ_s becomes less important in determining the pulse energy. In the extreme, the pulse energy approaches P_{cw}/f, and hence for two lasers with equal input (pump) powers, the pulse energies depend only on the relative CW laser efficiencies. In this regime, Nd:YAG offers shorter pulses, due mainly to its higher stimulated emission cross section, σ.

Higher CW power lasers lead to proportionately higher pulsed-mode output powers, and also to shorter laser pulse widths. Hence for communications systems, the benefit of higher-power lasers comes not only from basic transmitter power considerations, but also from the ability to limit background noise by using shorter communications slot widths. An understanding of the parameters affecting pulsed laser operation is important in designing communications lasers for maximum system performance.

References

[1] D. Sipes, "Highly Efficient Neodymium: Yttrium Aluminum Garnet Laser End Pumped by a Semiconductor Laser Array," *Applied Physics Letters*, 47(2), pp. 74–77, July 15, 1985.

[2] H. Hemmati, "Efficient Holmium: Yttrium Lithium Fluoride Laser Longitudinally Pumped by a Semiconductor Laser Array," *Applied Physics Letters*, 51(8), pp. 564–565, August 24, 1987.

[3] W. Koechner, *Solid-State Laser Engineering*, Chapter 8, New York: Springer-Verlag, pp. 397–408, 1976.

[4] G. D. Baldwin, "Output Power Calculations for a Continuously Pumped Q-Switched YAG:Nd^{+3} Laser," *IEEE Journal of Quantum Electronics*, vol. QE-7, no. 6, pp. 220–224, June 1971.

Table 1. Parameters for Nd:YAG and Ho:YLF lasers of Section IV

Parameters	Nd:YAG	Ho:YLF
P_{cw}, mW	100	100
τ_s, ms	0.23	12
λ, μm	1.064	2.06
V, cm^3	0.02	0.02
l, cm	1.0	1.0
L, cm	3.5	3.5
r_1	0.975	0.95
r_2	1.00	1.00
ξ	0.987	0.987
β, cm^{-1}	0.0023	0.0023
σ, cm^2	8.7×10^{-19}	1.0×10^{-19}

C-3

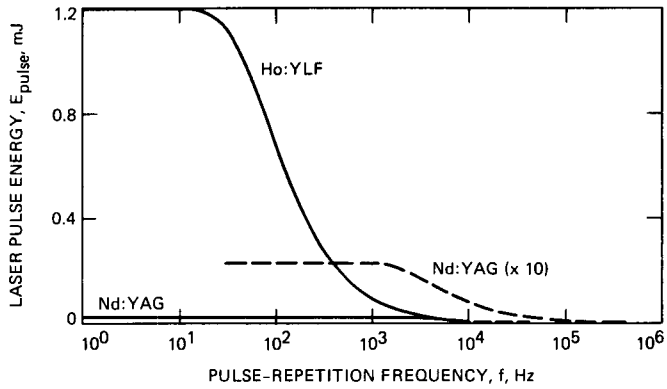


Fig. 1. Output energy per pulse, E_{pulse}, in mJ, versus pulse repetition frequency, f, in Hz, for Nd:YAG and Ho:YLF lasers described in Table 1. The curve marked Nd:YAG (X10) is the Nd:YAG curve multiplied by a factor of 10 for clarity.

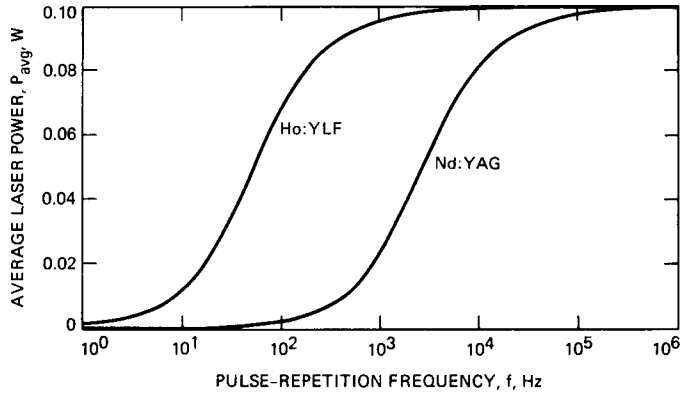


Fig. 2. Average output power, P_{avg}, versus pulse repetition frequency, f, for lasers of Table 1.

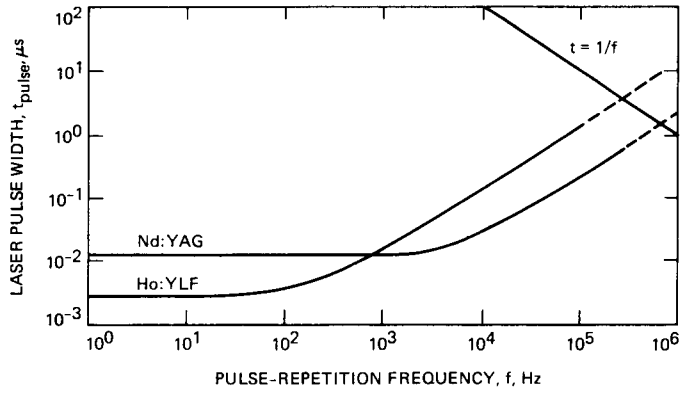


Fig. 3. Laser pulse width, t_{pulse}, versus pulse repetition frequency, f, for lasers of Table 1.

Calculations of Laser Cavity Dumping for Optical Communications

D. L. Robinson and M. D. Rayman
Communications Systems Research Section

For deep-space pulse-position modulation (PPM) optical communication links using Nd:YAG lasers, two types of laser transmitter modulation techniques are available for efficiently producing laser pulses over a broad range of repetition rates: Q-switching and cavity dumping. The desired modulation scheme is dependent on the required pulse repetition frequency and link parameters. These two techniques are discussed, theoretical and numerical calculations of the internal energy of the laser cavity in cavity dumping are described, and an example of cavity dumping is applied to a link for a proposed experiment package on Cassini.

I. Introduction

A link-analysis approach is a standard aspect of the development and design of a communications system. It is essential to have confidence that the component performances assumed in these link calculations are realizable. Because some of the key components in optical communications are still in the development phase, it is necessary to use theoretical analyses to support the performance assumptions made in the link studies.

One of these key components is the laser transmitter. The laser most likely to be used is a neodymium-doped yttrium aluminum garnet (Nd:YAG) crystal end-pumped by laser diodes. In contrast to flash-lamp pumping, laser diodes can provide pump light at one of the atomic resonant absorption bands of the Nd^{3+} ion to improve pumping efficiency. (It is the rare-Earth ion Nd^{3+} in the Nd:YAG that lases. The YAG is simply the host matrix.) By mode-matching the pump light into the laser cavity, the absorption of pump photons is made to occur exactly where the lasing takes place, and the absorption length is greater than in the common side-pumping geometry. With this architecture, overall electrical-to-optical efficiencies in excess of 10% have been demonstrated [1].

In order to implement deep-space optical communications, the extremely energy-efficient pulse-position modulation (PPM) scheme will be used. This modulation format puts severe demands on the performance of the laser transmitter, and it is very important to verify that the required performance that has been assumed in link calculations is achievable. In this article we report on progress in our understanding of the behavior of a modulated laser used for deep-space communications. Two regimes of modulation, Q-switching and cavity dumping, are discussed, and a study of a laser performing in the latter mode follows. Although this study does not model cavity dumping completely, it does provide valuable insight into the process. We evaluate the buildup of energy prior to

the emission of a signal pulse. This is a necessary step to insure that the energy required for deep-space optical communications is available in the laser cavity using realistic system parameters. The cavity-dumping analysis here is applied to an example of an optical communications link from Cassini during its interplanetary cruise. A detailed analysis of Q-switching will be presented in a future report.

II. Modulation Techniques and Applications

Two of the common techniques for achieving high-peak-power pulses from the Nd:YAG lasers are Q-switching and cavity dumping. In Q-switching, the energy is stored in the atomic population inversion by keeping the Q of the cavity too low to support laser oscillation. This is accomplished with the use of an element in the cavity whose loss can be controlled. Atoms are pumped to the upper state, but in the absence of stimulated emission, the upper-state population will be greater than in the equilibrium condition achieved when lasing occurs. When the Q is increased (by reducing the loss), the energy in the atoms is immediately available, and the stimulated-emission rate becomes large. A high-energy pulse then depletes the upper level, and lasing temporarily ceases. If the Q is reduced at that point, the pump energy will again begin accumulating population in the upper state. Q-switching has an upper limit imposed by the finite time required to repump the population inversion and by the cavity-field buildup time [2]. A pulse-repetition frequency (PRF) on the order of 50 kHz is the maximum value that can provide high-peak-power pulses from Q-switched Nd:YAG.

For pulse rates much higher than 50 kHz, the technique of cavity dumping is preferred. Although cavity dumping can be extremely efficient at frequencies of many megahertz, it is less efficient at low pulse-repetition rates. As PRFs increase, the choice between Q-switching and cavity dumping will depend on specific laser design parameters and link requirements. The optimal transition point will be understood after further study. In cavity dumping, instead of storing the energy in atoms, the energy is stored in the photon field of the cavity. The output-coupling strength is varied so that the energy in the cavity is extracted when it is needed. The laser is kept above threshold during the entire process.

Both of these modulation techniques have application to deep-space optical communications. Examples of specific optical links between a planetary spacecraft and Earth-based receivers will illustrate this. We have proposed to include an optical-communications package on Cassini. Second in the series of Mariner Mark II spacecraft, it will be targeted for Saturn orbit and will release a probe into the atmosphere of Titan. Currently, launch is expected in 1996. Although the prime communications system will use radio-frequency technology, there may be an opportunity to include an optical communications experiment package to prove out its technology, increase the data return rate, and perform a number of "light science" experiments which take advantage of the onboard laser, telescope, and other optical components.

One configuration of the Cassini optical package uses a 30-cm telescope for the transmit/receive antenna. A frequency-doubled Nd:YAG laser with an average power of 1 W would serve as the transmitter. Transmitting to a 10-m Earth-based receiver under clear skies [3], this system could return over 115 kb/sec from 9 AU. This includes Saturn being in the field of view of the receiver, and the calculated link margin is 3 dB [4]. To achieve this impressive performance, a PPM alphabet size of M=256 is used, and the width of each slot is 10 nsec. With the use of coding, the bit error rate is 10^{-5}.

Because PPM with M=256 transmits 8 bits per pulse, a data rate of 115 kb/sec requires 14,375 pulses per second. The duty cycle is obviously quite low, the laser being on for a total of only about 144 μsec each second. The dead time between the 256-slot words is 67 μsec. This mode of operation is comfortably in the Q-switch regime. With an average laser power of 1 W, each of the pulses has a peak power of almost 7 kW.

During interplanetary cruise, Cassini may be used to demonstrate much higher data rates. Using 256-ary PPM with 10-nsec slot widths and about 2.6-μsec dead time between words, the optical communications package could return 1.54 Mb/sec from 5 AU (the distance of Jupiter) with a 3-dB margin. This does not assume Jupiter to be in the field of view. To transmit 1.54 Mb/sec with M=256 PPM, the laser is required to emit 192.5 kilopulses per second. To maintain 1 W average power, each pulse requires a peak power of 519 W. Based on our present understanding, maximally efficient performance at this PRF necessitates the use of cavity dumping.

These examples illustrate the importance of both Q-switching and cavity dumping for deep-space optical communications. Detailed understanding of laser performance under both operating conditions is essential. In the following section, we present calculations of a laser using cavity dumping to achieve the higher Cassini data rate from 5 AU.

III. Analysis and Calculations of Cavity Dumping

It is important for us to understand the details of the behavior of a Nd:YAG laser operating in a cavity-dumping mode in order to make accurate predictions of its performance

and design an efficient system. Following the work of Chesler and Maydan [5], we can calculate the approximate performance of a Nd:YAG laser on Cassini at 5 AU as discussed above. These calculations describe the population inversion and internal field of the laser during buildup in preparation for emitting an output pulse. This initial approach to modeling cavity dumping does not include frequency doubling or the output pulse generation and its characteristics. But we shall see that it does allow us to determine and verify some important aspects of the laser performance. Chesler and Maydan begin with the rate equations for a continuously pumped laser:

$$\frac{dN}{dt} = R - \Gamma N - \beta FN \tag{1a}$$

and

$$\frac{dF}{dt} = \beta FN - (\epsilon + T)F \tag{1b}$$

In these equations, N is the number of atoms in the upper laser level; F is the number of coherent photons in the cavity; t is time; R is the number of atoms pumped up per second; Γ is the spontaneous-decay rate of the upper laser level; and βFN is the number of atoms per second undergoing stimulated emission. The stimulated emission coefficient β may be expressed as $c\sigma/AL$, where σ is the laser transition cross section, A is the cross-sectional area of the laser beam in the Nd:YAG rod, and L is the optical length of the cavity. $\epsilon \equiv c\Delta/2L$ is the reciprocal of the cavity decay time (not including losses from intentional output coupling), where Δ is the round-trip fractional inherent cavity loss. Similarly, $T \equiv c\alpha/2L$ is the reciprocal of the cavity decay time (including only intentional output coupling), where α is the fractional output coupling. During the buildup phase of the cavity dumping cycle, $\alpha = 0$. In order to extract a pulse, in the ideal case, the value of α would be changed to 1, thus allowing 100% of the stored energy to be emitted in a pulse. In reality some losses will be incurred in this process, but this analysis considers only the internal energy of the laser cavity.

These two equations can be understood by considering the physical processes involved in laser physics. Equation (1a) describes the time dependence of the atomic population inversion. The inversion is increased by atoms being pumped up to the upper laser level by the pump source, and it is diminished by both spontaneous and stimulated emission. The latter effect provides a positive contribution to the photon field in the cavity, and that is reflected in the first term on the right side of Eq. (1b). This equation describes the time dependence of the number of photons in the field of the cavity. The second term in that equation reflects the loss of photons through inherent and intentional losses in the cavity.

For a given cavity design, T will be fixed. It can be shown by maximizing the output power that the optimum cw values for N and F, given fixed T, are $N_0 = \epsilon\phi^{1/2}/\beta$ and $F_0 = \Gamma(\phi^{1/2} - 1)/\beta$. The parameter ϕ is defined to be $R\beta/\Gamma\epsilon$, which is the ratio of the pumping rate to the threshold pumping rate. Of course, in storing and dumping the energy in the cavity, the interest is in the deviations from the cw performance. Thus, we introduce n and f to describe these deviations, and we have $N = N_0 + nN_0$ and $F = F_0 + fF_0$.

Chesler and Maydan make a number of reasonable approximations to arrive at expressions for these deviations. One of the key assumptions is that the duration of an output pulse is short compared to the buildup time between pulses. In our example, this is seen to be an excellent assumption, since the pulse duration of 10 nsec is less than 0.4% of the minimum time between pulses. The approximate solutions are found to be

$$n = \frac{\gamma\Gamma}{\epsilon}\left(\frac{1}{\gamma} - \frac{1}{2} + \frac{s}{\tau} - \frac{e^{\gamma s/\tau}}{e^{\gamma} - 1}\right)$$

and

$$f = \frac{\gamma e^{\gamma s/\tau}}{e^{\gamma} - 1} - 1$$

where $s \equiv t\epsilon$, or the time in units of the cavity decay time; τ/ϵ is the time between pulses, during which the field intensity accumulates; and $\gamma \equiv \tau(\phi^{1/2} - 1)$.

With these expressions, we can calculate the evolution of the upper-state population and the optical field for cavity dumping in the regime of validity for these solutions. Because the development began with rate equations, the results do not apply when the number of coherent photons in the cavity is reduced to the order of one. At this level, the statistics of spontaneous emission control the buildup of the field, and the rate-equation approach is not appropriate.

Our interest now is in finding N/N_0 and F/F_0. We consider the case of a cavity with inherent loss $\Delta = 0.03$ and a length $L = 20$ cm. These combine to give a cavity decay time of $\epsilon = 44$ nsec. (Recall that ϵ does not include intentional output coupling. By increasing the output coupling, α, when it is time to emit the energy, 10-nsec pulses can be achieved. The technique used in this report to examine cavity dumping does not allow us to study the output pulse.) To calculate the parameter β, we use $\sigma \approx 5.75 \times 10^{-23}$ m^2 for Nd:YAG [6], and $A = 3.14 \times 10^{-6}$ m^2. Thus we find $\beta = 2.74 \times 10^{-8}$ Hz. The fluorescence lifetime of the upper state in the Nd:YAG laser line is 230 μsec, so $\Gamma = 4350$ Hz.

Using the output coupling which comes from the optimum cw values for N and F as outlined above, we can calculate a pumping rate which guarantees an average power of 1.0 W. This turns out to correspond to a pumping rate above threshold of $\phi = 4.9$. We know this is achievable, since this value of ϕ is less than that previously demonstrated for diode pumping of Nd:YAG lasers [7].

With these values, we determine n and f and thus N/N_0 and F/F_0 as functions of time. The results of these calculations are shown in Figs. 1 and 2.

At time 0 in both figures, the system is beginning just after a pulse has been produced. The pump energy is building up population in the upper level of the laser line and contributing to the field energy. When the field energy passes the cw optimum value of $F = F_0$ (Fig. 2), the rate of stimulated emission becomes large enough to begin reducing the population in the upper state. The upper-state population (N) begins to decline (Fig. 1), and it never varies significantly from the cw value. When the inversion decreases, energy is transferred into the optical field by stimulated emission until the designated time to dump the cavity. When it is time to produce a pulse, the output coupling (α) is changed, and the internal-field energy drops as it is emitted in the narrow pulse. The greatly reduced internal field causes a reduction in the stimulated-emission rate, so the population inversion begins to increase again and the entire cycle starts over.

From our calculations of the laser performance, we find that the cavity accumulates 5.17 μJ at the maximum. It is at that point that the pulse is produced by changing the output coupling. Although the approach used here does not address the dynamics of the output signal, if we assume that all of this available energy is emitted in a 10-nsec pulse, it produces a peak power of 517 W. This is within less than 1% of the values derived from the Cassini link calculation and is achieved with the laser component values we have used.

IV. Conclusions

Within the limitations of this initial approach to understanding cavity dumping, we can see that the performance assumed for the laser transmitter in the optical link calculations is justified. Realistic laser parameters with an achievable pumping rate will lead to production of the stored energy needed for the Cassini link from Jupiter.

A detailed understanding of the laser operation during cavity dumping is crucial to the design of a laser capable of providing the signals needed for the pulse-position modulation to be used in our optical-communications system. The approximate solutions used here provide a starting point for that understanding, but more needs to be done. One of the assumptions of the derivation is that the dumping is periodic. Of course, since the transmitted information is contained in the time during which the pulse is transmitted, varying times between pulses must be considered. This would allow the field energy and population-inversion energy to continue to evolve for different lengths of time between signals. A more detailed analysis of this factor would reveal exactly how it affects the uniformity of the output pulses. In addition, a study of pulsed pumping would be necessary in order to insure that the stored energy is maximum just before the output coupling is raised to release that energy as an output signal. Further, to achieve still higher data rates, operation in a regime where the time between pulses is not large compared to the pulse width is required, as has been assumed here. For Cassini at Mars range, data transmission of 20 Mb/sec is planned. To achieve that rate with 10-nsec pulses will require a dead time of only 40 nsec and an alphabet size of 16. An analysis of the performance of the laser transmitter under these conditions requires use of the exact solution. Such an analysis should include the actual extraction of the pulse to reveal its characteristics in detail. A careful comparison of Q-switching and cavity dumping in the PRF range where they overlap will allow the determination of the preferred scheme of modulation under different link scenarios.

References

[1] J. Berger, D. F. Welch, D. R. Scifres, W. Streifer, and P.S. Cross, "High power, high efficient neodymium:yttrium aluminum garnet laser end pumped by a laser diode array," *Appl. Phys. Lett.*, 51, pp. 1212-1214, 1987.

[2] R. B. Chesler, M. A. Karr, and J. E. Geusic, "An Experimental and Theoretical Study of High Repetition Rate Q-Switched Nd:YA1G Lasers," *Proceedings of the IEEE*, 58, pp. 1899-1914, 1970.

[3] E. L. Kerr, "Strawman Optical Reception Development Antenna (SORDA)," *TDA Progress Report 42-93*, Jet Propulsion Laboratory, Pasadena, California, pp. 97-110, May 15, 1988.

[4] W. K. Marshall and B. D. Burk, "Received Optical Power Calculations for Optical Communications Link Performance Analysis," *TDA Progress Report 42-87*, Jet Propulsion Laboratory, Pasadena, California, pp. 32-40, November 15, 1986.

[5] R. B. Chesler and D. Maydan, "Calculation of Nd:YA1G Cavity Dumping," *J. Appl. Phys.*, 42, pp. 1028-1030, 1971.

[6] W. Koechner, *Solid-State Laser Engineering*, New York: Springer-Verlag, 1976.

[7] Donald L. Sipes, Jr., "Nd:YAG End Pumped by Semiconductor Laser Arrays for Free Space Optical Communications," *IEEE Military Communications Conference*, Boston, MA, pp. 104-108, October 20-23, 1985.

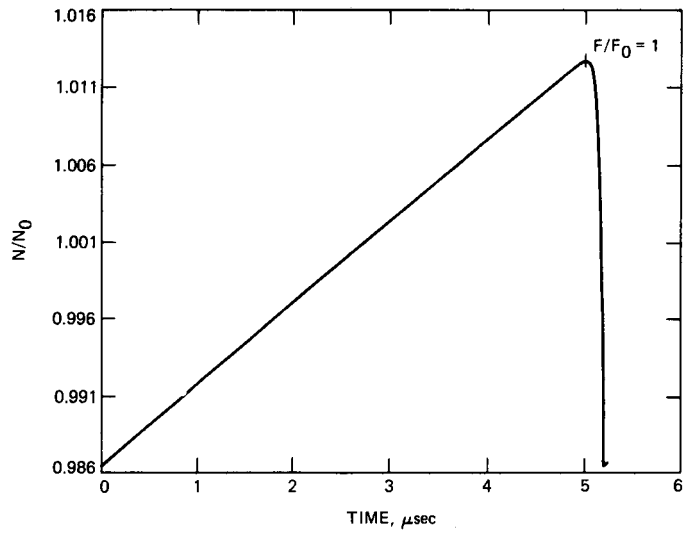


Fig. 1. Normalized population inversion as a function of time.

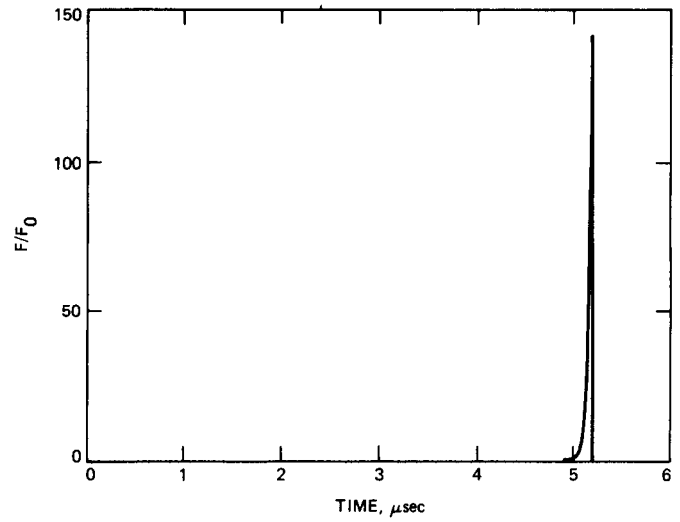


Fig. 2. Normalized field as a function of time.

An Integral Sunshade for Optical Reception Antennas

E. L. Kerr
Communications Systems Research Section

Optical reception antennas (telescopes) must be capable of receiving communications even when the deep-space laser source is located within a small angle of the Sun (small solar elongation). Direct sunlight must not be allowed to shine on the primary reflector of an optical reception antenna, because too much light would be scattered into the signal detectors. A conventional sunshade that does not obstruct the antenna aperture would have to be about five times longer than its diameter in order to receive optical communications at a solar elongation of 12 degrees without interference. Such a long sunshade could not be accommodated within the dome of any existing large-aperture astronomical facility, and providing a new dome large enough would be prohibitively expensive. It is also desirable to reduce the amount of energy a space-based large-aperture optical reception facility would expend orienting a structure with such a sizable moment of inertia.

Since a large-aperture optical reception antenna will probably have a hexagonally segmented primary reflector, a sunshade consisting of hexagonal tubes can be mounted in alignment with the segmentation without producing any additional geometric obstruction. The tubes can be extended downward toward the primary reflector, until they reach the envelope of the focused beam to the secondary reflector. If the optical reception antenna is ground-based, the other ends of the tubes may be trimmed so that both the sunshade and the antenna will fit within a sphere whose diameter is only six-fifths the diameter of the primary reflector. If the segmentation involves four rings of hexagons with the central segment absent from the primary, then this sunshade is useful when solar elongations are as small as 12 degrees. Additional vanes can be inserted in the hexagonal tubes to permit operation at 6 or 3 degrees. The structure of the sunshade is very strong and can be used to support the secondary reflector instead of an independent support.

An analysis of the duration and recurrence of solar-conjunction communications outages (caused when a deep-space probe near an outer planet appears to be closer to the Sun than a given minimum solar elongation), and the design equations for the integral sunshade are appended.

I. The Need for Sunshading

Direct-detection optical communication at visible wavelengths from laser sources on deep-space probes requires that background interference be reduced to acceptable levels. Background from natural sources is usually incoherent and can be reduced substantially by narrowband filtering. Additional immunity to interference is achieved by sending an optical pulse only during one time slot of a series of time slots. After filtering, the remaining background level must be low enough to ensure an acceptably small probability that the background count in any empty time slot will be less than the background plus signal count in the signal time slot.

For optical communication to a deep-space probe near an inner planet, the background interference may be so high that heterodyne detection techniques are required. These reduce background by using a post-detection filter whose bandwidth is much narrower than the bandwidth of any pre-detection optical filter. The spatial coherence of the signal must be preserved, however, which certainly requires good optics and may preclude reception through the Earth's atmosphere.

In studies of typical missions to outer planets it has been shown that the background is acceptably small for direct-detection optical communications even when the sunlit planet fills a substantial field of view behind the spacecraft ([1], Appendix A). The sunlight to be excluded by the sunshade is then direct sunlight scattered within the reception antenna (telescope) when the planet is near conjunction, i.e., at small Sun-Earth-probe angles (small solar elongations).

A. Scattering from Rough Reflectors

Recommended plans for the Optical Reception Development Antenna [2] call for use of a hexagonally segmented primary reflector made up of light-weight, composite panels having a root-mean-square surface roughness of $2\,\mu$m.[1] It is anticipated that sunlight directly incident on such a surface would produce intolerable scattering into the detectors, no matter what internal sunshades were used. Therefore, a primary sunshade must be provided, capable of shading the primary reflector from direct sunlight whenever the antenna is used at more than the design minimum solar elongation.

B. Thermal Effects on Visible Reception Antennas

Sunlight will not heat sunshades to incandescence; therefore reradiation from the sunshade will not be a problem for visible reception (even though it is a problem for infrared and millimeter-wave telescopes). Thermal distortion of the structure and convection currents or heat extraction difficulties are perennial problems deserving further study.

C. Communications Outages Near Solar Conjunction

The orbits of most of the planets lie close to the same plane, the plane of the ecliptic, which is the plane of the apparent path of the Sun through the sky and also the plane of the Earth's orbit. This means that the planets appear to approach the Sun as they are viewed from Earth. The times when the planets are close to the Sun are called conjunctions. The duration τ_d of a conjunction depends on the periods of revolution t of the Earth and T of the outer planet, and on the design minimum solar elongation E (the minimum solar elongation for communications is the maximum solar elongation of the conjunction that causes the communications outage). The period of recurrence τ_r of the conjunction depends on t and T.

Communications with a probe on a mission to an outer planet will be blocked whenever the planet appears to come too close to the Sun, as in Fig. 1. Table 1 shows the duration of outages for various limiting solar elongations, and the period of recurrence, for the outer planets. The formulas on which the table was based are derived in Appendix A.

The inclinations i of the orbital planes of each of the outer planets with respect to the Earth's orbital plane are also given in Table 1. The line of intersection of the orbital planes of a pair of planets is called the line of nodes. Only Pluto's orbit is highly inclined. This means that Pluto approaches a close conjunction only during the times when the Earth and Pluto are close to opposite ends of their line of nodes. At other times Pluto appears to pass at a variable angle (as much as 17 degrees) north or south of the Sun. The other outer planets move in orbital planes too close to the Earth's to help much in relieving the communications interference encountered near solar conjunction.

II. Disadvantages of Conventional Sunshades

The usual primary lightshade of a telescope is an internally blackened tube extending from the primary reflector to a short distance beyond the primary focus. This provides shading for the secondary reflector also, in a Cassegrain or Newtonian arrangement. Other internal baffles may be added. If a sunshade is needed, it is usually added as an extension of the primary lightshade beyond the primary focus. Sometimes this extension is cut down to a sun visor in order to reduce weight, though that requires the operation of orienting the telescope axially, relative to the Sun.

[1] P. N. Swanson, *A Lightweight Low Cost Large Deployable Reflector (LDR)*, JPL Publication D-2283 (internal document), Jet Propulsion Laboratory, Pasadena, California, pp. 5-1–5-6, June 1985.

The chief disadvantage of an extended primary lightshade is the length required in order to look at targets when the solar elongation is small, without allowing light to strike the primary reflector. The length required is $D \cot E$. If the telescope diameter D is 10 m and E = 12 degrees, the length is 47 m, which makes a very unwieldy telescope.

A series of slats or flat plates may be inserted within the tube, such that the normal to the plates is perpendicular to the line of sight. If the slats divide the diameter into n spaces of equal thickness between them, then the overall required sunshade length is divided by n. The number n cannot be made very large, however, for two reasons. First, the slats must be made of some material having a finite thickness, and the sunshade cannot be kept in perfect alignment with the line of sight. This means that the slats will obstruct the view to some extent, contributing some fraction to the opacity of the telescope. Second, the slats will introduce additional diffraction and spread the image of the deep-space laser source. The field stop will then have to be opened to capture a reasonable fraction of the incoming signal power. If a planet or another extended object is in the background, the background level will increase as the field stop is opened and the performance of the communications link will be degraded.

A. End-Mounted Sunshades

The largest telescopes, used for astronomy, do not have sunshades associated with them. Astronomers use the Earth as a natural sunshade by doing their observing at night. This is necessary because the natural objects they look at emit incoherent radiation which is too faint to be separated from the scattering in the daytime blue sky.

The dome is the most expensive component of an observatory building, and the cost increases faster than the square of the diameter. (The log-log graph in Fig. 2 has a best-fitting slope of 2.02 for standard, electrically-driven, hemispherical domes from 3 to 11 m in diameter. The 37-m Keck dome is custom-made, more than hemispherical, and has special drives and sensors for precise positioning. The slope from the largest standard dome to the Keck dome is 3.81.) For this reason, among others, the dome is usually made only large enough to clear the swing sphere swept out by the motion of the telescope.

1. **Ground-based antennas.** If an astronomical observatory with a large-diameter telescope were converted or rented for use as a ground-based optical reception station, an end-mounted sunshade of a reasonable length could not be accommodated within the dome, for the reasons stated above.

2. **Space-based antennas.** A space-based optical reception antenna could conceivably be sunshaded by a tube (or visor or even a flat plate) whose length was about five times the diameter of the antenna. Ingenious methods could be devised to transport such a structure to space, erect, and assemble it. However, the moment of inertia would be very large. A great deal of energy would be expended in orienting the telescope and sunshade while tracking a deep-space probe.

B. Externally Mounted Sunshades for Ground-Based Reception

Since most large telescopes are protected by a dome, it would be possible to mount a long tube externally on the dome. Alignment of the tube with the telescope is necessary but is not required to be very precise. A small computer could easily control the azimuth of the dome and the elevation angle of the sunshade as well as the azimuth and elevation of the telescope, when tracking an object and compensating for the rotation of the Earth.

However, most observatory sites are on mountain peaks where they are subject to occasional high winds. Table Mountain Observatory, for example, reports clocking winds at 90 m/sec (200 mph), which their domes survive. An externally mounted sunshade would add considerable wind load to the dome. The dome would have to be strengthened for operation in moderate winds, and the sunshade would have to be stowed securely whenever high winds or inclement weather was anticipated.

III. Solution: The Integral Sunshade

An integral sunshade for optical reception antennas is proposed to overcome the disadvantages detailed above. Figures 3 and 4 are photographs of a glue-and-paper model of the integral sunshade. The sunshade consists of a bundle of closely packed hexagonal tubes, aligned with the antenna line of sight, forming a structure that supports the secondary reflector. (This makes the sunshade an integral part of the telescope structure.) The outer edges of the outermost tubes extend around the primary reflector and form the primary sunshade. Inside, the tubes are cut off just short enough to provide clearance for the focused beam to the secondary reflector. The opposite ends are trimmed in the form of a spherical cap, to fit within the swing sphere of the telescope.

A plan view of the reflector is shown in Fig. 5. The axial hexagons are those that straddle the x-axis. The central hexagon is numbered (0,0). The other hexagons are numbered first by their ring number (starting with the innermost) and then by their sequence number within the ring (starting with the axial hexagon). The numbered hexagons all have different surface figures in order to fit together into a parabolic reflector. The points of individual hexagons have been lettered A, B, C, D, E,

and F, counterclockwise starting with the point closest to the 60-degree line. This lettering is illustrated for hexagons (1,1) and (3,3).

The other ring hexagons are symmetrical with respect to rotations of 60 degrees. The axial hexagon on the outermost ring, (4,1) in Fig. 5, is called a corner hexagon.

Figure 6 shows the integral sunshade concept as it would be for a Cassegrain optical reception antenna having a 10-m, $f/0.5$, hexagonally segmented, four-ring, primary reflector. An x-z cross-section and a y-z cross-section are shown, each covering only the positive half of the x- or y-axis. The reflector lies at the bottom of, and is tangent to, a 12-m-diameter swing sphere. The secondary reflector is the same size as the absent central hexagonal panel of the primary reflector.

Another baffle surrounds the hole in the primary left by the absent central hexagon. This baffle consists of the frustum of a six-sided pyramid cut off by the intersection of two planes to form each edge. One plane is determined by the edge of the central hexagon and the primary focal point. The other is determined by the edge of the secondary reflector and the secondary focal point. The pyramid is illustrated in each cross-section of Fig. 6 by a line from the primary reflector slanting inward.

Within the pyramid is a conical (or cylindrical) baffle designed to capture rays that enter the pyramid after passing through the innermost ring tube. This baffle is illustrated in each cross-section of Fig. 6 by a line from the center of the primary reflector slanting outward.

A. Design Premises

1. Sunlight may not be allowed to shine anywhere on the primary reflector. Premise 1 could be violated, of course, by looking at a deep-space probe at less than the allowed solar elongation. Sunlight would not flood the entire primary until the solar elongation was equal to half the solar subtense. A small amount of sunlight scattering from the primary might be tolerable, depending on the parameters of the optical communications link. However, Premise 1 is used to define the minimum solar elongation for normal operation.

2. A ray of sunlight is considered to be stopped when incident, however obliquely, on the blackened surface of any part of the sunshade. Premise 2 does not require the existence of a perfect absorber with no forward scattering. It only means that the absorption is adequate to reduce the interference caused by the remaining forward-scattered sunlight to tolerable levels.

3. The sunshade parts and reflecting surfaces are considered to be infinitesimally thin. Premise 3 actually renders the design conservative. In practice it is expected that there will be gaps of about 2 cm or so between the panels of the primary reflector. The walls of the hexagonal tubes will not have to be nearly so thick to form a very strong structure. The absorption of obliquely incident rays on their surfaces may therefore be improved by adding ridges or ring baffles consisting of thin plates cut to fit perpendicularly within the tubes, having a hexagonal hole punched in them the same size as the reflecting panel below. A light ray incident on the tube wall just above such a ring would experience two geometrical reflections, one from the tube wall followed by one from the ring baffle, such that the ray would actually be reflected back parallel to itself, as illustrated in Fig. 7.

Ring baffles effectively reduce the chords across the tubes without introducing any additional geometrical obscuration of the telescope aperture beyond that produced by the segmentation. Reduction of the chords without changing the lengths of the tubes means that the sunshade could be used at solar elongations slightly less than the minimum.

4. The truss supporting the primary reflector and the optics behind it can all be fitted in the space between the primary reflector and the swing sphere. If Premise 4 cannot be fulfilled in fact, the dome will have to be made with a somewhat larger clearance.

The design equations are set forth in Appendix B.

B. Design for Operation Within a Minimal Dome

During the initial conception of this design it was observed that the top ends of the hexagonal tubes follow a curve that parallels, to some extent, the cone of the focused beam from the primary to the secondary. For an $f/0.5$ primary the parallelism is good when the diameter of the swing sphere is six-fifths of the diameter of the primary. This allows operation within a dome that fits very closely over the optical reception antenna. Such a dome and sunshaded telescope are illustrated in Fig. 8. The dome opening is far larger than that of conventional domes with meridional shutters, however, and the dome must be much more than hemispherical if the telescope is to be able to look horizontally or down to some minimal elevation angle.

The two shortest sets of tubes are then the tubes on the innermost ring, and those on the corners of the outermost ring. Rays entering along the longest chords within the innermost tubes are stopped by the pyramid, however. The minimum solar elongation is equal to the arcsine of the ratio of the

longest chord to the shortest distance through the corner tube on the outermost ring from top to bottom perimeters. The longest chord is between the points of the tube. The shortest distance goes from either of the top outermost points to the opposite bottom inner point.

In the design shown in Figs. 3 through 6 the minimum solar elongation E is 12.44 degrees. Allowing for ring baffles 1 cm wide reduces E to 11.96 degrees.

The minimum solar elongation may be cut approximately in half by inserting a set of plates or vanes between the points of each tube, so the cross section resembles an asterisk inscribed within a hexagon (Fig. 9a). The plates would run the length of the tube, and would be cut off at the ends to fit the primary focused beam and the swing sphere, just as the tubes are. This effectively subdivides the hexagonally segmented aperture into equilateral triangles. This time additional obscuration and diffraction are introduced.

Reduction of the minimum solar elongation to one quarter can be accomplished by subdividing each of the equilateral triangles again with plates. The cross section resembles a six-pointed star superimposed over the asterisk and inscribed within a hexagon (Fig. 9b). This structure would be stable without the circumscribing hexagon (unlike the asterisk structure). The tubes could be designed with channels running the length of each point, and the six-point-star-and-asterisk vanes could be inserted in each tube whenever it was necessary to track an object at close solar conjunction. The vanes could be removed whenever operations did not require looking closer than 12 degrees of the Sun to eliminate the obscuration and diffraction the vanes cause.

C. Application to a Space-Based Reception Antenna

A space-based antenna would not have to swing within a prescribed sphere. However, the integral sunshade has a moment of inertia that is considerably smaller than that of an open tube providing similar sunshading, and the center of gravity is much closer to the primary focal point. These factors favor use of the integral sunshade for space-based optical reception antennas operating in the visible region of the spectrum.

The structure of the sunshade is very strong and rigid. It may be used to mount the secondary reflector at a fixed distance from the primary. The mass and added diffraction of a secondary-reflector support spider are eliminated.

D. Summary of Design Advantages

A very compact, manageable sunshade is provided. No geometrical obscuration not introduced already by segmentation of the primary reflector is added. The amount of diffraction added by the sunshade is very small.

The swing sphere for the entire system is only slightly larger than the sphere needed to swing the primary reflector. No wind loads are added to the dome. Dismounting and stowing an external sunshade for anticipated inclement weather are not required.

The center of the swing sphere is placed very close to the primary focal point. (A small adjustment of the focal length of the primary reflector would make the two points coincide, if that were desirable.) The design provides a rigid structure to support the secondary reflector above the primary. The mass and added diffraction of a secondary support spider are eliminated.

E. Possible Extensions of the Design

The minimum solar elongation is set by skew rays through the tubes on the corners of the outermost ring. It is possible to reduce the minimum solar elongation by adding some small additional baffles.

One method would simply cap a portion of the outer top edge of the corner tube whenever the telescope is used at a small solar elongation. The area of the effective aperture would be reduced only slightly, and a somewhat smaller solar elongation would be allowed.

Another method would add some short vertical baffles arranged along radial lines at the inner points of the bottoms of the corner tubes on the outer ring. The entering collimated beam from the deep-space probe and the focused beam to the secondary reflector would be obstructed by these vertical radial baffles only to the extent of their finite thickness. However, they could be made long enough to come close to the surface of the primary reflector (to within a suitable clearance), and could intercept the skew rays that had previously limited the minimum solar elongation. The minimum solar elongation might therefore be reduced to a new limit imposed by skew rays in another set of tubes, either the nearest neighbors of the corner tubes on the outermost ring, or the tubes that form the next-to-innermost ring.

F. Areas Requiring Further Study

1. Use of radial baffles within the focused beam region between reflectors. Small radial extensions of the baffles on the outermost-ring corner tubes have been mentioned already.

Inspection of Fig. 5 shows that some of the segmentation lines are radial, on the odd-numbered rings beginning with the

innermost. Since the integral sunshade forms a very strong structure, and radial plates introduce an additional geometrical obstruction proportional only to their thickness, a designer might consider extending the radial walls of the tubes downwards to tie together the integral sunshade and the primary reflector support truss. Very little additional sunshading would be provided, but a more rigid overall structure would be obtained. Thermal analysis would have to show that the advantage in rigidity would not be offset by the thermal distortions caused by nonuniform heating of the sunshade.

2. **Thermal problems to be overcome.** Absorption heating will occur on one side only of the tubes, with the depth of penetration dependent on the location of the tubes relative to the Sun. Due allowance must be made for thermal distortion of the structure, and possible dislocation of the secondary reflector.

a. Ground-based antennas. Convection currents within the tubes will arise if a large temperature difference exists between opposite walls. These currents produce fluctuations in the density and refractive index of the air, and would blur the image of the deep-space laser source at the focal point of the system. The threshold for the onset of convection, i.e., the ratio of the temperature difference between opposite walls to the absolute temperature, is inversely proportional to the cube of the distance between the walls [3]. The walls must therefore be highly thermally conducting, in order to reduce the temperature difference between opposite walls within a tube as much as possible. The central tube can be left open at the top, with a thermal shield over the secondary reflector at the bottom, in order to ensure uniform heating of its walls. This also suggests that the outer surfaces of the sunshade should be blackened, contrary to the normal practice of making the primary sunshade white on the outside.

The tubes on the side closest to the Sun will be penetrated to a greater depth, and through a larger projected aperture, than the tubes on the side farthest from the Sun. When the deep-space probe is seen above the Sun this situation can lead to gravity-driven circulation of air that will enter the lower tubes, pass between the reflectors, and exit through the upper tubes. As long as the flow is laminar the blurring of the image will be minimal. However, the dynamics of this process require study.

Forced outward convection of filtered air through all the tubes may be capable of providing necessary cooling, preventing unstable or turbulent convection, and keeping the optics clean.

b. Space-based antennas. Lack of convection will eliminate blurring of the optical image but may require introduction of heat pipes or other means of heat extraction.

3. **Use of the sunshade instead of a dome.** Microwave radio antennas are often used without domes. It may be possible to provide tube caps and weatherization that would eliminate the need for a dome over the optical reception antenna, at a substantial savings in cost.

4. **Mass reduction and deployability of a space-based integral sunshade.** Deployment of a low-mass integral sunshade in space represents a solvable construction challenge, especially if the integral sunshade is used as proposed to support the secondary reflector in relation to the primary reflector instead of a support spider.

IV. Conclusions and Recommendations

A novel kind of sunshade has been proposed for large-aperture hexagonally segmented optical reception antennas, to permit optical communication even when the deep-space laser source is as close to the Sun as 12 degrees. Inserts in the tubes of the sunshade would permit operations at solar elongations as small as 6 or 3 degrees, at a slight reduction in effective aperture area and a small increase in diffraction spreading of the source image.

The compactness of the sunshade effects a substantial cost savings when the optical reception antenna is ground-based and housed under a dome. The inner diameter of the dome can be almost as small as six-fifths of the aperture diameter. A space-based optical reception antenna would use much less energy to orient this sunshade than it would orienting a conventional sunshade of comparable functionality, and the mass and added diffraction of a secondary-reflector support spider are eliminated.

A few design issues remain for investigation, such as the thermal distortion, convection currents with ground-based antennas, and heat extraction for space-based antennas.

References

[1] J. R. Lesh and D. L. Robinson, "A Cost-Performance Model for Ground-Based Optical Communications Receiving Telescopes," *TDA Progress Report 42-87*, vol. July–September 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 56–64, November 15, 1986.

[2] E. L. Kerr, "Strawman Optical Reception Development Antenna (SORDA)," *TDA Progress Report 42-93*, vol. January-March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 97–110, May 15, 1988.

[3] J. W. Strutt, "On Convection Currents in a Horizontal Layer of Fluid, When the Higher Temperature is on the Under Side," *Philosophical Magazine*, vol. 32, pp. 529–546, 1916. Reprinted in *Scientific Papers, Cambridge, 1920*, New York: Dover Publications, vol. 6, pp. 432–446, 1964.

Table 1. Recurrence and duration of solar conjunctions

Planet	T, sec	τ_r, yr, d	Elongation, deg				i, deg
			12	6	3	1	
			τ_d, d	τ_d, d	τ_d, d	τ_d, d	
Mars	59355300	2 49	86	43	22	7	1.850
Jupiter	374320000	1 34	32	16	8	3	1.309
Saturn	929604000	1 13	28	14	7	2	2.493
Uranus	2651140000	1 4	26	13	6	2	0.773
Neptune	5200270000	1 2	25	13	6	2	1.779
Pluto	7837350000	1 1	25	13	6	2	17.146

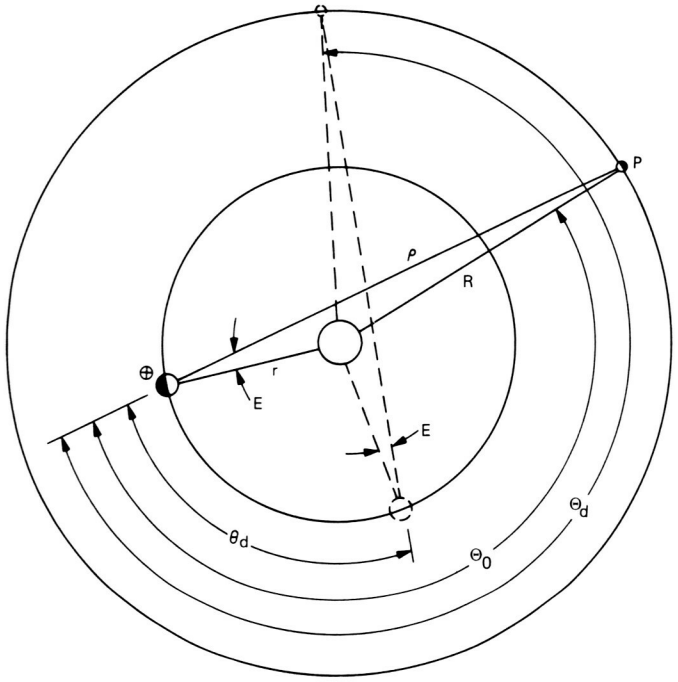


Fig. 1. Configuration of the Earth (⊕) and an outer planet P when approaching (solid lines) and leaving (broken lines) solar conjunction within solar elongation *E*.

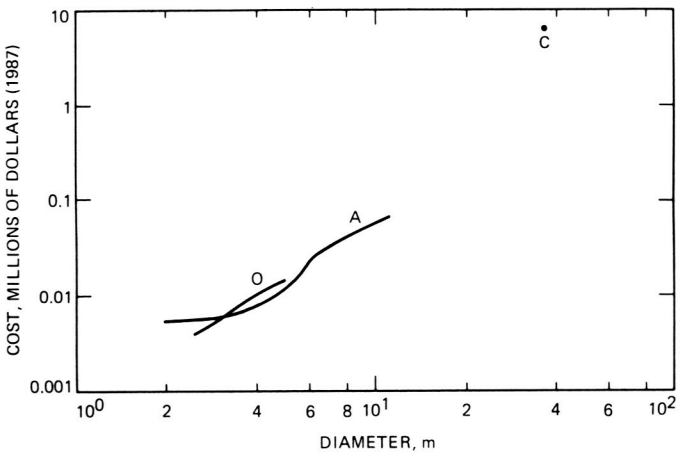


Fig. 2. Logarithm of dome cost versus dome diameter. Data supplied by manufacturers: A = Ash-Dome, C = Coast Steel, O = Observa-Dome Laboratories.

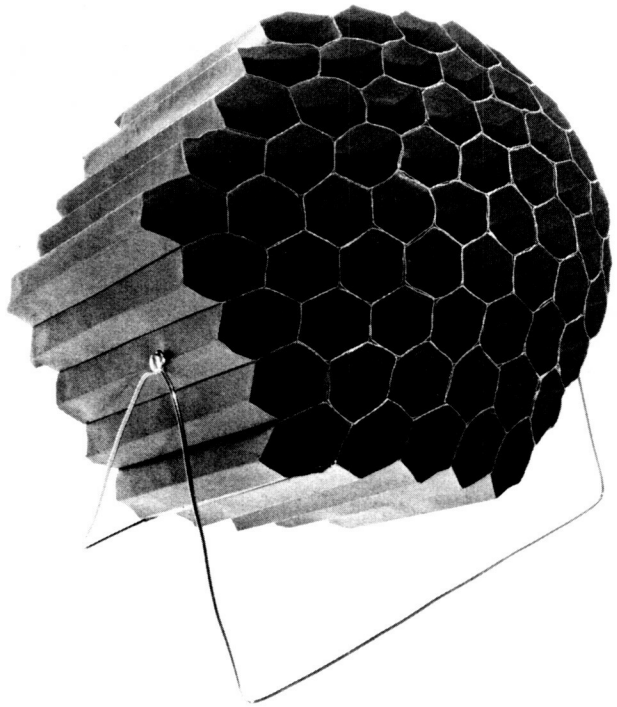


Fig. 3. Overview of a glue-and paper model of an integral sunshade. The telescope looks through the sunshade from behind and below it, in this view. The hexagonal tubes are trimmed to a spherical shape.

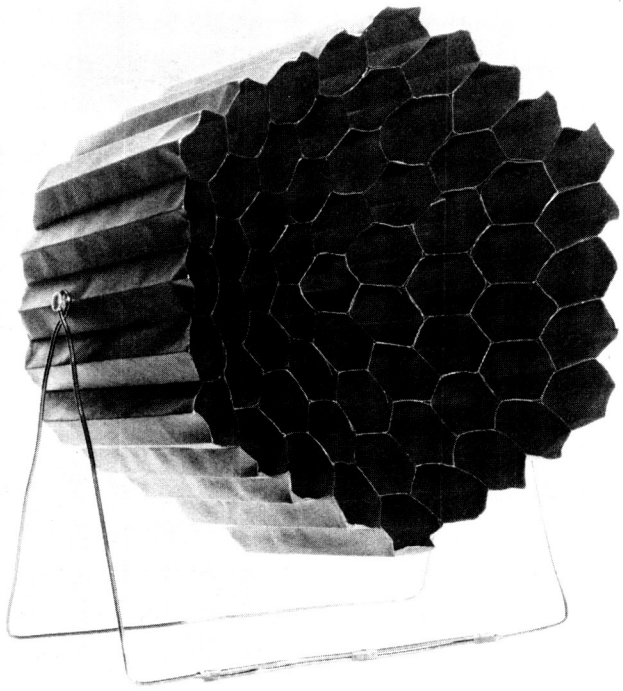


Fig. 4. Underside view of an integral sunshade model. The hexagonal tubes are trimmed to form a six-sided pyramid, with the secondary reflector to be mounted at the apex and the primary reflector at the base.

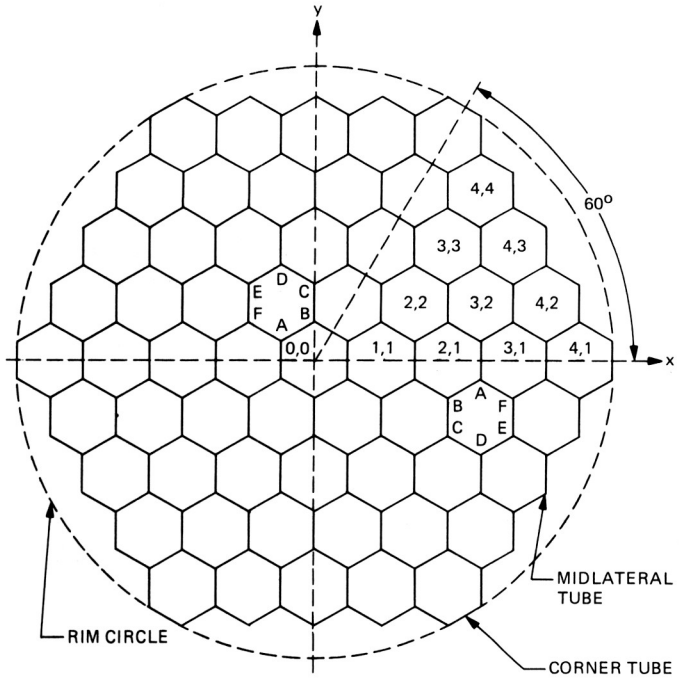


Fig. 5. Plan view of a hexagonally segmented reflector.

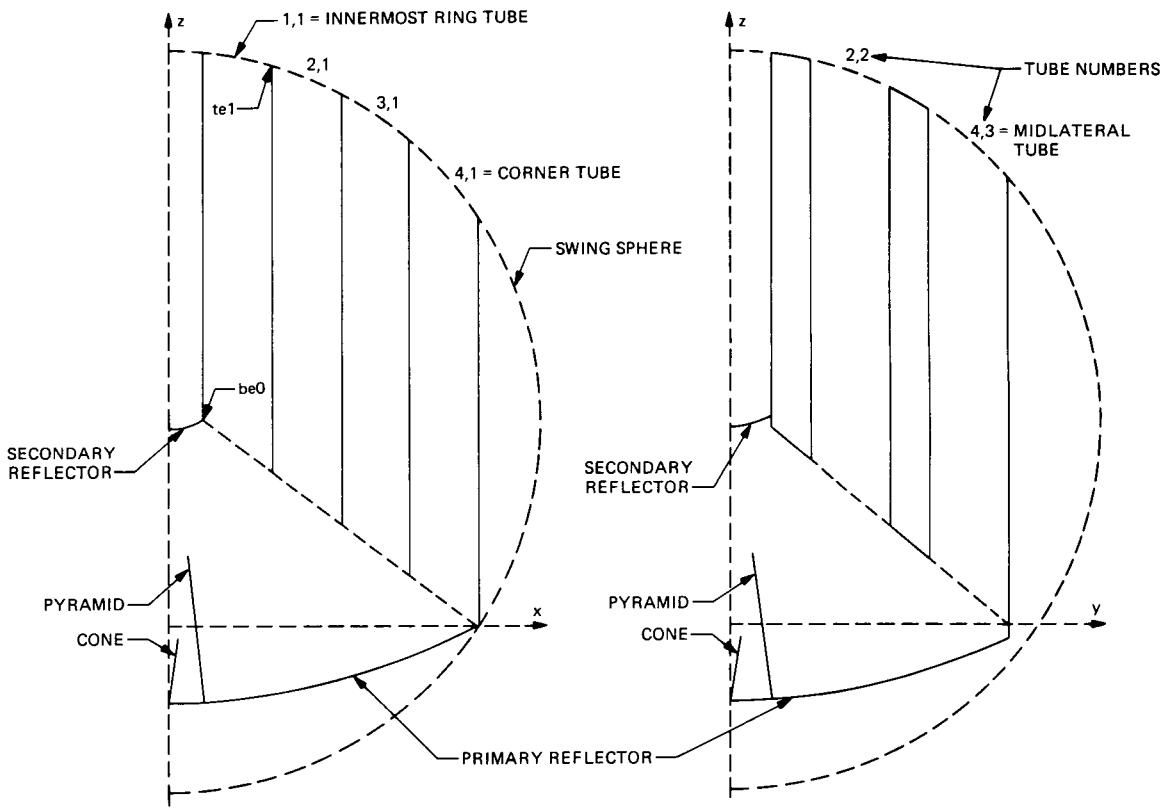


Fig. 6. Cross-sectional views of an integral sunshade on a Cassegrain optical reception antenna within a swing sphere.

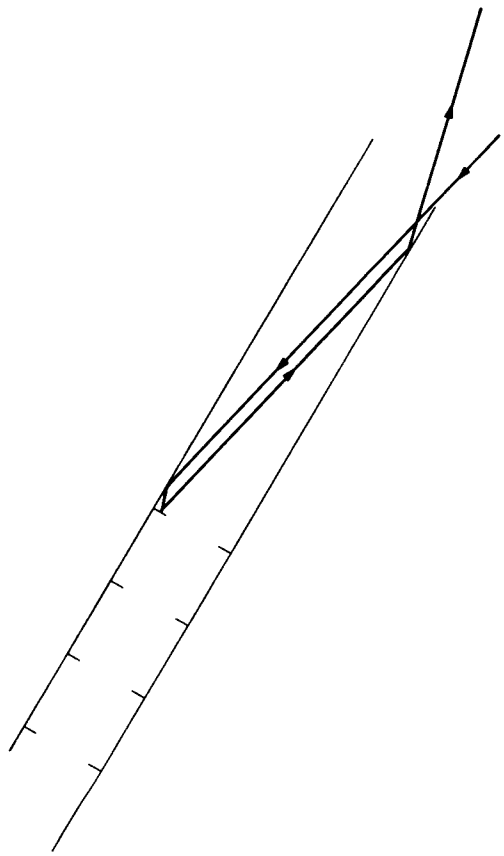


Fig. 7. Geometrical reflection of rays from a tube wall and ring baffles.

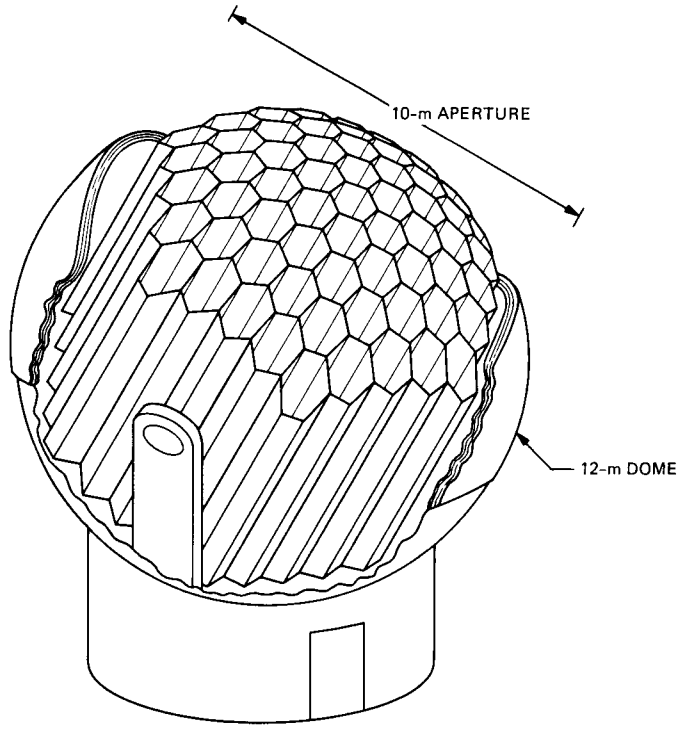


Fig. 8. An optical reception antenna and integral sunshade fitting within a spherical dome whose diameter is six-fifths the aperture diameter.

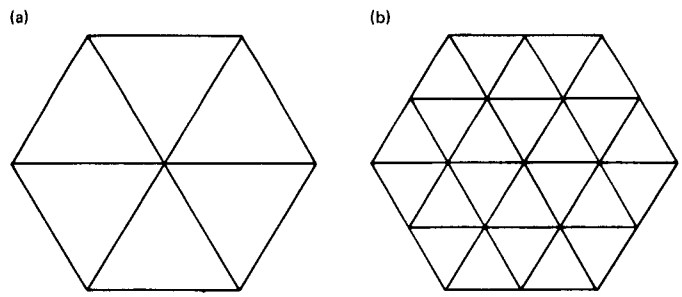


Fig. 9. Cross sections of (a) asterisk vanes within a haxagonal tube and (b) six-point-star vanes superimposed over asterisk vanes.

Appendix A
Duration and Recurrence of Outer-Planet Conjunctions

The orbits of the Earth and an outer planet P of the solar system are shown in Fig. 1 as viewed from above the north pole of the Sun. The respective periods of revolution t and T are related to the respective orbital semimajor diameters r and R by Kepler's law, $r^3/t^2 = R^3/T^2$, which shows that the outer planet moves at a slower angular rate $2\pi/T$ than the Earth's angular rate $2\pi/t$. The zero of angular measurement is designated as the Earth's position at the time that the outer planet is seen from Earth at an elongation angle E east of the Sun, when the planet is approaching conjunction. The analysis will be simplified by approximating the orbits with circles lying in the same plane and concentric on the Sun. The distance between the Earth and the outer planet is ρ, and the angular position of the outer planet is Θ_0. Trigonometric relationships for the solid-line Earth–Sun–outer-planet triangle yield

$$\frac{\sin(2\pi - \Theta_0)}{\rho} = \frac{\sin E}{R}, \quad \rho^2 = r^2 + R^2 - 2rR\cos(2\pi - \Theta_0)$$

Squaring the first relationship, substituting for ρ^2 from the second, and replacing $\sin^2(2\pi - \Theta_0)$ with $1 - \cos^2(2\pi - \Theta_0)$ leads to a quadratic equation whose solution is

$$\Theta_0 = 2\pi - \cos^{-1}\left[\frac{r}{R}\sin^2 E \pm \cos E\sqrt{1 - \frac{r^2}{R^2}\sin^2 E}\right]$$

The lower sign corresponds to the solution sought. The upper sign leads to a solution that becomes degenerate if one considers conjunction of a planet in the same orbit as the Earth, so $R = r$. (The upper-sign solution would correspond to a position coincident with the Earth, and the ratio $(\sin E)/R$ would be equal to the ambiguous form 0/0.) The ratio r/R may be replaced with $(t/T)^{2/3}$ in order to work only with the planetary revolution periods.

The orbital position of the Earth at any time τ is $\theta = 2\pi\tau/t$, and that of the outer planet is $\Theta = \Theta_0 + 2\pi\tau/T$. At the time τ_d when the outer planet has passed conjunction and is seen at an angle E west of the Sun (i.e., at the end of the time that the outer planet is seen within an elongation E of the Sun), the Earth has reached the position θ_d, the outer planet has reached the position Θ_d, and the configuration is represented by the broken-line triangle. The solid- and broken-line triangles are congruent since the radii and the elongation angles are equal, so the two obtuse angles are equal,

$$2\pi - \Theta_0 = \Theta_d - \theta_d = \Theta_0 + 2\pi\tau_d\left(\frac{1}{T} - \frac{1}{t}\right)$$

The solution for the duration is

$$\tau_d = \frac{\left(\frac{\Theta_0}{\pi} - 1\right)}{\left(\frac{1}{t} - \frac{1}{T}\right)}$$

The next occurrence of an epoch of conjunction begins at τ_r, the first time the difference between the orbital positions $(\Theta - \theta)$ modulo 2π is again equal to the original difference Θ_0. The outer planet moves more slowly so the difference becomes negative and 2π will have to be added. This gives

$$\Theta_0 + \frac{2\pi\tau_r}{T} - \frac{2\pi\tau_r}{t} + 2\pi = \Theta_0$$

$$\tau_r = \frac{1}{\left(\frac{1}{t} - \frac{1}{T}\right)}$$

Appendix B
Design Equations

The integral sunshade for an optical reception antenna was designed on the basis of a parameterized model of the antenna. The antenna is assumed to consist of a parabolic primary reflector and a secondary reflector in the form of part of the upper sheet of a hyperboloid of revolution. The primary reflector is hexagonally segmented and rests in the bottom of the swing sphere. The secondary reflector is of the same size as the absent central hexagon of the primary reflector. The secondary reflector is positioned where it will be filled by the focused beam from the primary reflector, and the secondary focus is centered on the projected surface of the primary reflector.

The following definitions, parameter values, and equations were used.

Origin: Center of primary reflector rim circle

Positive z axis: Along the line of sight

Positive x axis: Through the center of a corner hexagon

R = 6 m = radius of telescope swing sphere

D = 10 m = aperture diameter

n = 4 = number of rings

$z_c = \sqrt{R^2 - (D^2/4)}$ = 3.317 m; center of swing sphere is at $(0,0,z_c)$

ψ = polar angle for swing sphere

$x = R\sin(\psi)$ = x-coordinate of swing sphere

$z = R\cos(\psi) + z_c$ = z-coordinate of swing sphere

$w = D/(2n + 1)$ = 1.111 m = width across flats of a hexagonal segment

$s = w/\sqrt{3}$ = 0.642 m = side length of hexagonal segment

f = 0.5 = f number or focal ratio of primary reflector

Paraboloidal primary reflector surface equation

$$z = \frac{(x^2 + y^2)}{4fD} - \frac{D}{16f}$$

$D/(16f)$ = 1.250 m = reflector concavity

$fD = F$ = 5.000 m = distance from reflector center to primary focus

$z_{Fp} = fD - [D/(16f)]$ = 3.750 m; primary focus is located at $(0,0,z_{Fp})$

h = 0.833 m = axial step height between tube edges at reflector ends

$$h = 2\frac{fD - [D/(16f)]}{2n + 1}$$

$f_{fDw} = (fD/2) - fw + [w/(16f)]$ = 2.083 m = distance involved in determining the secondary reflector

$fs = (f^2D^2/4) + (w^2/4) + f_{fDw}^2$ = 10.899 m^2 = area factor involved in determining the secondary reflector

$a^2 = \left(f_S - \sqrt{f_S^2 - f^2D^2f_{fDw}^2}\right)/2$ = 3.846 m^2 = hyperboloidal secondary major parameter squared

$b^2 = c^2 - a^2 = (f^2D^2/4) - a^2$ = 2.404 m^2 = hyperboloidal secondary minor parameter squared

$z_h = (fD/2) - [D/(16f)]$ = 1.250 m = center of hyperboloidal secondary

$y_{omlp} = -[(3n/2) + 1]\,s$ = −4.490 m = outer point of outer midlateral hexagon on primary

$z_{omlp} = -[D/(16f)] + [y_{omlp}^2/(4fD)]$ = 0.242 m = outer point of outer midlateral hexagon on primary

$z = z_h + \sqrt{a^2 + (x^2 + y^2)(a^2/b^2)}$ = equation of hyperbolic secondary reflector

$z_{be0} = nh$ = 3.333 m = height at bottom of central tube edge

A coordinate system was established in the plane of the primary reflector to make the positions of most features of the hexagonal grid equal to an integral number of units. In some cases, however, half-units had to be counted.

$u = w/2$ = 0.555 m = 1 unit in the x-direction

$t = s/2$ = 0.321 m = 1 unit in the y-direction

The cutting of the lower ends of the tubes to clear the focused beam from the primary reflector forms a kind of "ceiling" over the primary. This ceiling is in the form of six planes, convergent at the primary focus, forming a six-sided pyramid. The base of the pyramid was chosen to rest on the outer points of the outer-ring point hexagons. Each plane was

then determined by the primary focal point and two other base points, of which the following two are typical:

x_{c1} = 9 (in units of u), y_{c1} = 1 (in units of t), z_{c1} = 0 m

x_{c2} = 5 (in units of u), y_{c2} = 13 (in units of t), z_{c2} = 0 m

The minimum elongation angle E is the smallest angle that allows a ray to penetrate the sunshade to the primary reflector surface. This ray is one that crosses from a low point on the top of one hexagonal tube to a high opposite point on the tube at the bottom, in a short tube, provided that afterwards the ray is incident on the primary reflector. The angle is calculated by giving the coordinates of the two opposite points of a tube, as follows, and trying various tubes until the largest angle E is found. In the following, subscript t refers to the top point, and b to the bottom point.

x_t = 9 (in units of u), y_t = −1 (in units of t)

$z_t = z_c + \sqrt{R^2 - x_t^2 u^2 - y_t^2 t^2}$ = coordinate (on sphere)

x_b = 7 (in units of u), y_b = 1 (in units of t)

z_b = coordinate (focused beam clearance)

$$= z_{Fp} - \frac{x_b[(z_{Fp} - z_{c1})y_{c2} - (z_{Fp} - z_{c2})y_{c1}] + y_b[x_{c1}(z_{Fp} - z_{c2}) - x_{c2}(z_{Fp} - z_{c1})]}{x_{c1}y_{c2} - x_{c2}y_{c1}}$$

E = minimum elongation angle

$$= \frac{180}{\pi} \tan^{-1} \frac{\sqrt{(x_t - x_b)^2 u^2 + (y_t - y_b)^2 t^2}}{z_t - z_b}$$

Each unique off-axis tube stands over an area completely within a sector defined by the positive x-axis and a radial line at 60 degrees from it. The axial tubes are split by the x-axis, with their positive halves standing over the sector just defined. The points of the tubes were lettered A, B, C, D, E, and F, in order counterclockwise starting with the point nearest the 60-degree sector line. The top and bottom coordinates for cutting the tubes were then found as the intersection of the vertical lines of the tubes (the projections of the hexagon points) with the "ceiling" or with the swing sphere.

i = number of the ring (center tube is 0)

j = number of hexagon in ring, from 1 on x-axis up to i

$$x_A = \begin{cases} 1 & \text{if } i = 0 \\ 2i - j & \text{otherwise} \end{cases}, \quad y_A = \begin{cases} 1 & \text{if } i = 0 \\ 3j - 4 & \text{otherwise} \end{cases}$$

$$x_B = \begin{cases} 1 & \text{if } i = 0 \\ 2i - j + 1 & \text{otherwise} \end{cases}, \quad y_B = \begin{cases} 1 & \text{if } i = 0 \\ 3j - 5 & \text{otherwise} \end{cases}$$

$$x_C = \begin{cases} 1 & \text{if } i = 0 \\ 2i - j + 2 & \text{otherwise} \end{cases}, \quad y_C = \begin{cases} 1 & \text{if } i = 0 \\ 3j - 4 & \text{otherwise} \end{cases}$$

$$x_D = \begin{cases} 1 & \text{if } i = 0 \\ 2i - j + 2 & \text{otherwise} \end{cases}, \quad y_D = \begin{cases} 1 & \text{if } i = 0 \\ 3j - 2 & \text{otherwise} \end{cases}$$

$$x_E = \begin{cases} 1 & \text{if } i = 0 \\ 2i - j + 1 & \text{otherwise} \end{cases}, \quad y_E = \begin{cases} 1 & \text{if } i = 0 \\ 3j - 1 & \text{otherwise} \end{cases}$$

$$x_F = \begin{cases} 1 & \text{if } i = 0 \\ 2i - j & \text{otherwise} \end{cases}, \quad y_F = \begin{cases} 1 & \text{if } i = 0 \\ 3j - 2 & \text{otherwise} \end{cases}$$

Let X stand for one of the letters A through F. Then

$$z_{top} = z_c + \sqrt{R^2 - x_X^2 u^2 - y_X^2 t^2}$$

$$z_{bottom} = z_{Fp} - \frac{x_X[(z_{Fp} - z_{c1})y_{c2} - (z_{Fp} - z_{c2})y_{c1}] + |y_X|[x_{c1}(z_{Fp} - z_{c2}) - x_{c2}(z_{Fp} - z_{c1})]}{x_{c1}y_{c2} - x_{c2}y_{c1}}$$

A six-sided frustum of a pyramid, placed around the aperture in the primary reflector, prevents the entrance of any rescattered stray light from the primary reflector into the detector. It also limits the view from the secondary focal point to the secondary reflector. The larger base of the frustum is the same size as a hexagonal segment. The height above the primary reflector rim plane is determined by the intersection of a ray from the secondary focal point to a midlateral point of the secondary reflector and a ray from the primary focal point to a midlateral point of the central hexagon on the primary reflector. The calculations involve a pyramid factor $f_{pyr} = F/[z_{be0} + D/(16f) + F] = 0.522$. The height above the primary reflector mirror rim plane is $z_{Fp} - Ff_{pyr} = 1.145$ m. At the top of the pyramid, the distance to the edge from the center is $wf_{pyr}/2 = 0.290$ m, and the distance to the point from the center is $-sf_{pyr} = 0.335$ m.

A hexagonal cone (or a hexagonal cylinder) from the secondary focal point around the beam from the secondary reflector may be added to capture rays that are rescattered within the structure, or that enter through the inner ring tubes at less than the minimum elongation but would strike the primary reflector surface within the central hexagon. The height and spreading of the cone are determined by the intersection of the ray from the secondary focal point to a midlateral point of the secondary reflector and a ray passing through the innermost ring tube from the top outer midlateral point to the bottom inner midlateral point. This gives the distance from the center to the top edge of the cone as

$$w\,\frac{z_{te1} - 3z_{be0} - [2D/(16f)]}{2z_{te1} + 2z_{be0} + [D/(4f)]} = 0.127 \text{ m}$$

and the height (relative to the primary reflector rim plane) as

$$z_{te1} + \left[\frac{z_{te1} - 3z_{be0} - [2D/(16f)]}{2z_{te1} + 2z_{be0} + [D/(4f)]} - 3/2\right](z_{te1} - z_{be0})$$

$$= -0.199 \text{ m}$$

Shutters and Slats for the Integral Sunshade of an Optical Reception Antenna

E. L. Kerr and C. W. DeVore
Communications Systems Research Section

Optical reception antennas used at a small Sun-Earth-probe angle (small solar elongation E) require sunshading to prevent intolerable scattering of light from the surface of the primary mirror. An integral sunshade consisting of hexagonal tubes aligned with the segmentation of a large mirror has been proposed for use down to E = 12 degrees. For smaller angles, asterisk-shaped vanes inserted into the length of the hexagonal tubes would allow operation down to about 6 degrees with a fixed obscuration of 3.6 percent. Here we investigate two alternative methods of extending the usefulness of the integral sunshade to smaller angles by adding either variable-area shutters to block the tube corners that admit off-axis sunlight or by inserting slats (partial vanes) down the full length of some tubes. Slats are effective for most operations down to 6 degrees, and obscure only 1.2 percent. For E between 10.75 and 12 degrees, shutters cause even less obscuration.

I. Introduction

Deep-space-to-earth optical communication will require the development of a large-aperture ground-based reception antenna. Such an antenna, SORDA, is described in [1]. To receive data from a deep-space probe during the daylight hours, it is essential to shade the antenna primary mirror from all sunlight. Various sunshade designs have been considered. The best so far, the integral sunshade, is described in [2]. The integral sunshade segments the aperture with a bundle of long hexagonal tubes. During solar conjunction, when the space probe appears to be close to the Sun, the integral sunshade blocks sunlight incidence on the primary mirror as long as the angle E seen from the Earth between the Sun and the probe (the SEP angle or the solar elongation) is greater than 12 degrees.

When the solar elongation is less than 12 degrees, the integral sunshade would admit "chinks" or oddly-shaped patches of sunlight on the primary mirror, as in Fig. 1(a). The chinks would grow as the solar elongation was reduced; see Fig. 1(b). In [2] it was proposed that vanes (six flat plates arranged in cross section like an asterisk) be inserted in the hexagonal tubes of the sunshade to permit operation within 6 degrees of the Sun. The finite thickness of the vanes would obstruct the antenna aperture (that is, reduce its effective collecting area). If the width across the flat sides of the hexagonal tubes is 1.11 m and the effective vane thickness is 1 cm, the area reduction is 3.6 percent.

Herein, we compare two modified approaches to the sunshading problem. Since the chinks begin from sunlight leakage through the extreme corners of the hexagonal tubes, the corners may be blocked with appropriately shaped partial shutters, whose area can be increased as the solar elongation is reduced. These shutters would also obstruct the aperture, but

at first not as much as the vanes would. The distribution of sizes and the additional obstruction caused by the partial shutters are analyzed in this article. Alternatively, one may insert a slat down the length of each tube as the decreasing solar elongation exposes it to sunlight penetration. We will compare the effectiveness of the shutters to the slats.

II. Review of the Integral Sunshade

Sunshades used with small antennas are usually an extension of the primary lightshade between the objective and eyepiece, or between the secondary and primary mirrors. The closer the antenna is pointed toward the Sun, the longer the sunshade must be. End-mounted sunshades become impractical for large antennas. For instance, at $E = 12$ degrees, the sunshade length must be five times the aperture diameter. In one previous concept, the sunshade was to be mounted on the exterior of the antenna dome. Later it was observed that, by segmenting the sunshade in a hexagonal pattern similar to that proposed for the primary mirror, the sunshade may be shortened enough to mount on a fast-primary antenna and fit almost within the spherical volume swept out by the motion of the primary mirror itself. This configuration saves cost by minimizing the dome enlargement required to accommodate the sunshade. It eliminates the problem of having to remove and secure the sunshade during times of high winds. The integral sunshade can also provide a strong, rigid support structure for the secondary reflector, instead of the usual spider.

The hexagonal tubes extend at one end as close as possible to the envelope of the focused beam from the primary mirror, and to the reception station dome at the other end. If employed with a segmented primary mirror, the tubes would be aligned with the segmentation lines to minimize obscuration. The integral sunshade is short enough to fit within a dome whose diameter is 6/5 the diameter of the primary mirror.

The current design calls for a unit composed of sixty-one tubes. The width across the flats of each tube is 1.11 meters. The integral sunshade has the appearance of a honeycomb, concave in the shape of a pyramid on the bottom, and convex in the shape of a sphere on the top. The sunshade and antenna are supported at elevation pivot points on opposite sides of the sunshade by a yoke whose base may be turned in azimuth. Within each tube, along the sides, are ring baffles to intercept unabsorbed glare light reflected from the side of the tube at grazing incidence, and to redirect it back out the top of the tube. The ring baffles make the effective wall thickness of the integral sunshade about equal to 1 centimeter. The integral sunshade excludes sunlight from the primary mirror except when the solar elongation is less than 12 degrees.

III. Review of Solar Conjunctions

Most space probes, particularly those on missions to the planets, appear to approach and recede from the Sun because of the orbiting of the Earth and of the space probe. The times when this happens are called epochs of solar conjunction. The chief reason for tracking a probe near the Sun is that the probe may be near a planet that has periodic solar conjunctions. An outer planet will appear to approach the Sun from the east, pass close to it above or below or perhaps exactly behind it, and then recede to the west, over a period of several days. The path above, behind, or below the Sun depends on the inclination and orientation of the planet orbit relative to the Earth orbit. During the days of closest approach, or when the planet moves directly behind the Sun, there will be a communications outage. The duration of the outage may be reduced by reducing the minimum solar elongation at which optical communication is possible.

IV. Design Complications

Fitting the upper part of the sunshade to the swing sphere leads to some complications, which have been turned into opportunities for design economies. Because of the variations in the lengths of the tubes, the shading characteristics are not uniform. In general, the tubes nearest the Sun will begin to admit light at small elongation angles before the tubes farther away. This means that small shutters can be added on the corners of some tubes while others are left open, with little overall obstruction of the collecting area.

The exact shape and placement of the shutters, and the sizing of each one, depend not only on the solar elongation, but also on the direction of the line seen in the sky from the probe to the Sun. The line will appear to rotate slowly relative to the segmentation pattern as the antenna moves in azimuth and elevation to track the probe over a maximum of about ten hours of observing, from some minimum elevation angle at rising to the same angle at setting. The shutters would also have to move and change in size appropriately, or else be made a little larger than the absolute minimum necessary. The geometry of this problem is very complicated and still under study. Further investigation will also determine the exact conditions under which a slat will be effective in blocking solar penetration.

Operations and usage will affect the detailed design and implementation of the shutters. During any given day, the shutters may remain about the same size, but their shape and placement depend on the path taken by the planet during conjunction. The number of slats required would also be constant for a day, but some of them might have to be removed

and turned during a day (or else two slats in the form of an "X" would be required.) After the time of closest approach during the conjunction, the shutters or slats must all be switched to the other side of the antenna. The switching could be effected most easily by making the range of the elevation axis a full 180 degrees, and by rotating 180 degrees in azimuth also. However, the difficulties of mounting the mirror segments to maintain their alignment during a reversal of the gravity vector would probably limit the elevation range to about 110 degrees.

V. Methods of Reducing the Minimum Solar Elongation for Optical Reception

The first alternative would be to block a corner or side of each vulnerable tube at the outer end. A study of this approach has been undertaken. In the study, the integral sunshade structural plates were taken to be infinitesimally thin. The 1-centimeter thickness added by the ring baffles would allow operation about 0.5 degree closer to the Sun. Thus far, only the worst case has been analyzed. The Sun will penetrate the shade most easily if the Sun is shining at an angle across opposite vertices of the hexagons. It will begin doing so when E is less than 12 degrees. (In the best case, when the Sun is shining perpendicularly to opposite edges of the hexagons, sunlight penetration does not begin until E becomes less than about 10.5 degrees.) Since different tubes would require different amounts of extra shading, the size of the added shade in each of the tubes would vary. Figure 2 shows the sizes and shapes of the shutters in the case where $E = 9$ degrees with an attendant loss of 3.6 percent of the collection area. The case where $E = 6.5$ degrees is illustrated in Fig. 3. Here the loss of signal would be 26.7 percent. The amount of the remaining collecting area has also been calculated and graphed in Fig. 4. The collecting area drops off to 73.3 percent at an elongation angle of 7.049 degrees. (Elongation angles were calculated and shown in Fig. 4 for the infinitesimally thin-walled sunshade. Elsewhere in this report they have been estimated and reported to be about 0.5 degrees smaller because the ring baffles will be used.) Further study is necessary to analyze the problem of shading as the angle of the Sun in relation to the target changes. It appears, however, that only a small fraction of additional obscuration is required to permit uninterrupted observation for as long as ten hours, while the Sun angles change from early morning to late afternoon.

A second alternative calls for inserting a slat into each tube as shading is required. This will cause a 1.2-percent loss of signal to the corresponding mirror segment. As the solar elongation decreases from 12 degrees, the total receiver area loss for the telescope rises from 0 to 1.2 percent. This loss is much more acceptable than the larger losses mentioned earlier for shutters. However, slats are much larger than shutters and potentially more difficult to install or implement for automatic placement. They require reorientation (just as the shutters do) as the Sun angle changes.

As illustrated in Fig. 5, there is a range of angles from 10.75 to 12 degrees for which the shutters give slightly less obscuration than the slats. As the elongation diminishes from 10.75 degrees, the performance of the slats relative to the shutters increases dramatically.

VI. Conclusion

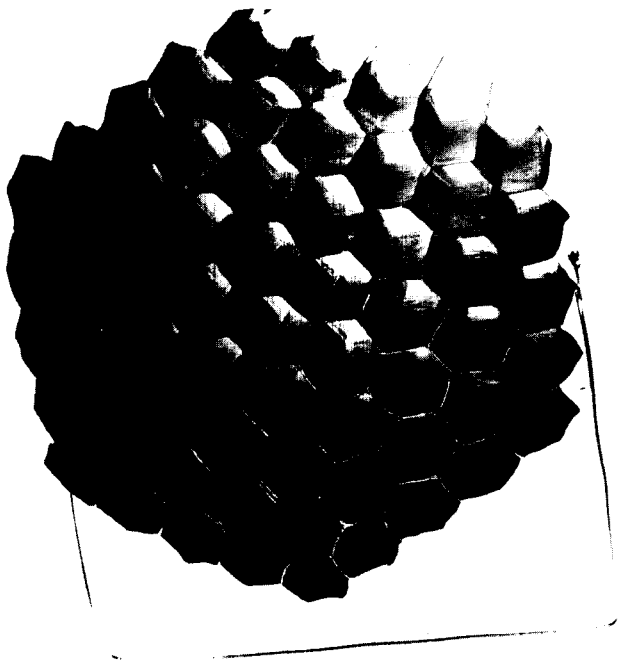
As currently conceived, without any additional shading, the integral sunshade will block unwanted solar interference for any solar elongation down to 12 degrees. By attaching variable-area partial shutters at the ends of some of the tubes, it will be possible to continue to receive optical communication from a space probe with a loss in signal power varying from 0 to 3.6 percent as the elongation is reduced from 12 degrees to 9 degrees. Slats inserted across the corners and along the length of the hexagonal tubes would cause an overall signal loss varying from 0 to 1.2 percent as the solar elongation varies from 12 degrees to 8 degrees. The signal loss with shutters is slightly less than the loss with slats for elongations from 12 degrees down to 10.5 degrees; for smaller angles the loss is decidedly greater. If the optical communications system performance is such that a greater signal loss can be accepted, it may be more convenient to use shutters instead of vanes even at elongations as small as 6.5 degrees. The shutters should be much easier to fabricate than the slats, and it should be much easier to actuate the shutters than to install the slats when needed. Both shutters and slats should be considered during the design of the optical reception antenna.

References

[1] E. L. Kerr, "Strawman Optical Reception Development Antenna (SORDA)," *TDA Progress Report 42-93*, Jet Propulsion Laboratory, Pasadena, California, pp. 97-110, May 15, 1988.

[2] E. L. Kerr, "An Integral Sunshade for Optical Reception Antennas," *TDA Progress Report 42-95*, this issue.

(a)



(b)

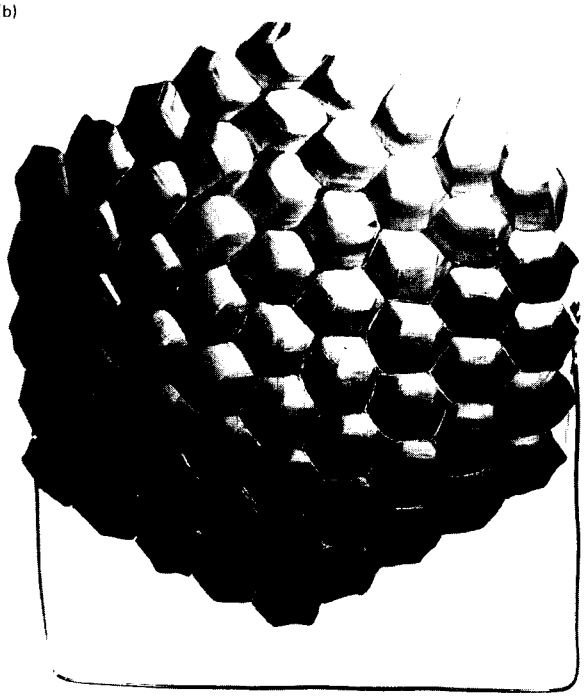


Fig. 1. A model of the integral sunshade. The primary mirror is to be mounted on the back, facing forward. (a) Light incident at this small angle relative to the sunshade axis is admitted by the tubes on the upper right that show patches of the background. (b) At this smaller angle more tubes show the background and admit more light.

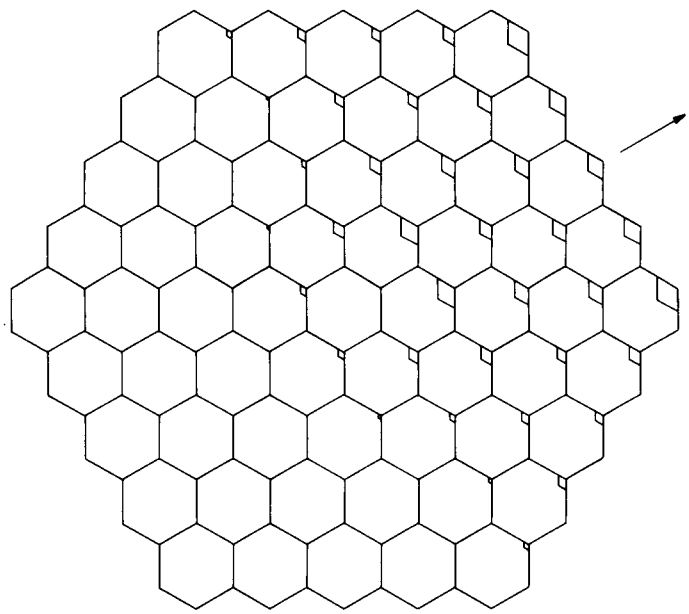


Fig. 2. Partial shutters as viewed along the length of the integral sunshade tubes, when the projected Sun vector (arrow) is directed parallel to the line joining the hexagon corners from lower left to upper right. This figure shows shading necessary for $E = 9$ degrees.

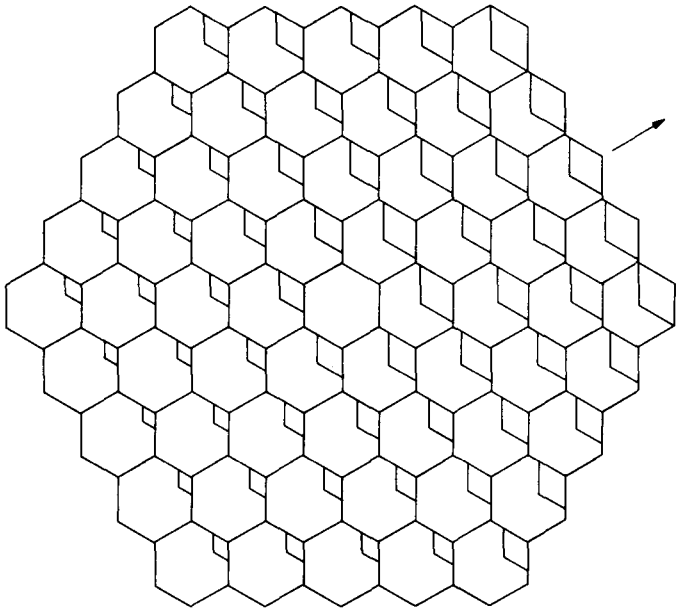


Fig. 3. Partial shutters when sunlight is incident along the same line as in Fig. 2. Here we see the shading necessary for E = 6.5 degrees, an extreme case in which some shutters block a third of the tube area.

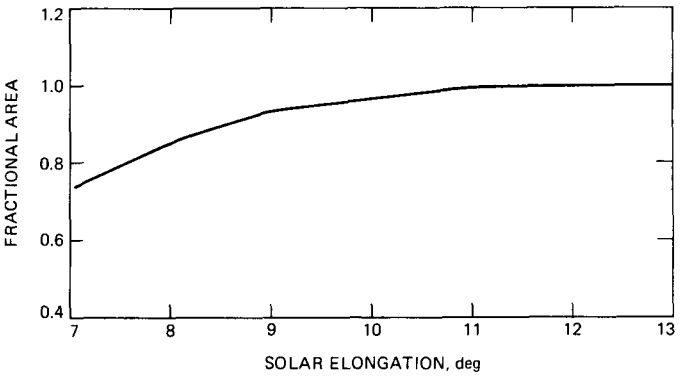


Fig. 4. Collecting area reduction as a function of solar elongation when minimal partial shutters are used with the worst sunlight incidence direction (when the sunlight vector as projected to the rim plane of the primary mirror is parallel to a line joining opposite vertices of a hexagon). Numerical values are based on the infinitesimally thin-walled integral sunshade model.

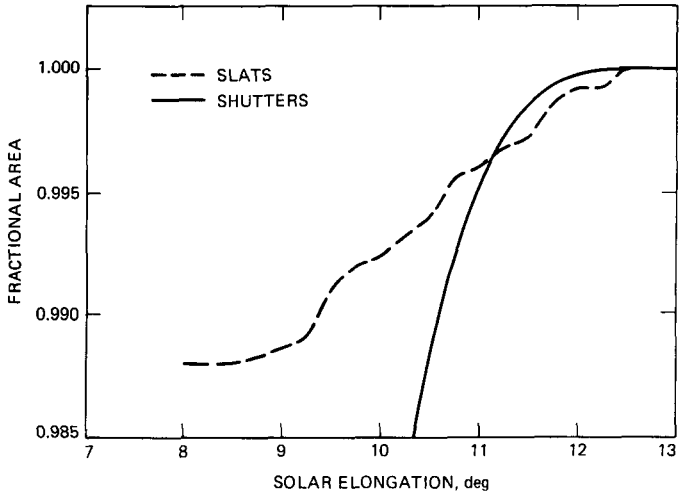


Fig. 5. Comparison of collecting area reduction for shutters and slats.

Effect of Earth Albedo Variation on the Performance of a Spatial Acquisition Subsystem Aboard a Planetary Spacecraft

C.-C. Chen
Communications Systems Research Section

The effect of Earth albedo variation on the pointing and tracking subsystem of a planetary optical communication package is analyzed. By studying the Cramer-Rao bound of the tracking error variance, it is shown that, when the Earth albedo is precisely known, the variance in spatial tracking error is inversely proportional to the total signal count. In contrast, a small uncertainty in the Earth albedo can result in an irreducible error in the tracking subsystem.

I. Introduction

Accurate spatial acquisition and tracking are critical for the operation of free-space optical communication systems. In order to maintain the signal power loss to within an acceptable level, tracking accuracy on the order of 1/10 to 1/20 of the transmitted beamwidth is generally required. For a system operating at 0.5-μm wavelength using a 30-cm aperture, the desired pointing accuracy will be on the order of 0.1 μrad. Since the angular resolution of the optical system is roughly equal to its transmitted beamwidth, the narrow pointing requirement implies that the spatial acquisition subsystem must derive a pointing reference to within 10 percent of the minimal spatial resolution. Because both the transmitter and the receiver are typically in motion, the tracking information must be derived within a time period during which the receiver may move a significant distance within the transmitter field-of-view. In some systems, the residual vibration due to the mechanical system will be much larger than the desired pointing accuracy. For these systems, the tracking information must be obtained at a rate that is higher than the vibration frequency so that the optical system can effectively compensate for vibration-induced error.

The performance of spatial tracking algorithms has been investigated by several authors [1], [2]. Most of these studies, however, were carried out with the assumption that a beacon signal is available as a pointing reference. These studies generally suggest that the effective pointing error decreases with increasing beacon strength or, effectively, with increasing integration time. In the latter case it is assumed that the beacon strength remains constant so that increasing the integration time, i.e., decreasing the tracking loop bandwidth, will result in an effective increase of the detector signal-to-noise ratio (SNR). In some cases, however, an active pointing beacon can be either undesirable or unfeasible. In these cases it may be desirable for the transmitter to use the a priori knowledge of the receiver to derive the pointing information. For instance, a spacecraft may use the sun-lit Earth as a pointing reference

and derive the actual receiver location using simple geometric rules. This scheme can be particularly attractive for a deep-space optical communication system for which the uplink beacon may require several kilowatts of radiated power.

The problem of deriving a pointing reference is equivalent to locating the image of the object on the receiver focal plane. For an extended object that can be resolved by the telescope, the optimal maximum-likelihood tracking algorithm has been derived, and its performance has been extensively investigated. The results generally state that, given the known source intensity distribution, the variance of tracking error will be inversely proportional to the detector SNR. Unfortunately, for most applications, the source brightness distribution is not precisely known. Consequently, there will be an error associated with the spatial tracking subsystem. The purpose of this report is to analyze the effect of uncertainties in source brightness on the spatial tracking subsystem.

The rest of this paper is organized as follows. Section II describes the conventional maximum-likelihood tracking algorithm for determining the pointing reference. The effect of uncertainties in source brightness will be analyzed in Section III by studying the Cramer-Rao bound on the tracking error variance. A typical planetary optical communication package using the sun-lit Earth as a pointing reference will then be described and the impact of Earth albedo variation on the tracking error will be studied in Section IV. Finally, the results from the study will be summarized in Section V.

II. Maximum Likelihood Spatial Tracking Algorithm

In this section the Maximum Likelihood (ML) algorithm for determining the angular position of the receiver is derived. It is assumed that the shape and orientation of the receiver is known. The problem of estimating the angular coordinate of the receiver is equivalent to estimating the location of the receiver image on the receiver focal plane. Without loss of generality, this is equivalent to locating a fixed reference point on the image. For simplicity, the reference point is chosen to be the geometric center r of the image. Consequently, the problem of spatial tracking can be reduced to the problem of estimating the geometric center given the detector photocount statistics and the prior knowledge of the source brightness distribution, $I_0(\rho)$, where ρ is the distance to the image center.

The problem of spatial estimating is complicated by the fact that the receiver has a finite spatial resolution. This is because the tracking detector, which is typically a charge-coupled device (CCD), has discrete spatial cells (pixels) that occupy a finite area. Each pixel output represents the total intensity of light impinged on the pixel area. Furthermore, because of the granularity of the optical signal, the output of the $(i,j)^{th}$ pixel k_{ij} will be a Poisson-distributed random variable with a mean $\lambda_{ij}(r)$ where

$$\lambda_{ij}(r) = \int_{A_{ij}} I_0(\rho - r)\, d^2\rho \tag{1}$$

In writing Eq. (1), $I_0(\rho - r)$ denotes the brightness distribution of the image where the geometric center is displaced by an amount r, and the integral is over the surface of the $(i,j)^{th}$ pixel.

The problem of deriving the transmitter pointing reference is therefore reduced to the problem of estimating the deviation r from a set of detector photocounts, $\{k_{ij}\}$. The maximum a posteriori (MAP) estimator [3] of the deviation $\hat{r}_{MAP}$ is given by

$$\hat{r}_{MAP} = \arg\left\{\max_r \{P(r|\{k_{ij}\})\}\right\} \tag{2}$$

If it is assumed that the prior probability distribution of r_0 is uniformly distributed, then the decision rule can be reduced to the following maximum-likelihood (ML) rule:

$$\hat{r} = \arg\left\{\max_r \{P(\{k_{ij}\}|r)\}\right\} = \arg\left\{\max_r \left\{\log \prod_{i=1}^{N}\prod_{j=1}^{N} \frac{\lambda_{ij}(r)^{k_{ij}}}{k_{ij}!} \times e^{-\lambda_{ij}(r)}\right\}\right\} \tag{3}$$

where it is assumed that k_{ij} are independent and Poisson-distributed with parameter $\lambda_{ij}(r)$, and observed that the logarithmic transformation does not change the maximum of a function. The ML decision rule can be further simplified by realizing that

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{ij}(r)$$

is proportional to the total power received by the receiver, and is therefore independent of the location of the geometric center r. Consequently, the ML decision rule can be written as

$$\hat{r} = \arg\left\{\max_r \left\{\sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij} \log \lambda_{ij}(r)\right\}\right\} \tag{4}$$

The performance of the estimator can be estimated by using the Cramer-Rao lower bound (CRLB) [1]

$$\mathrm{Var}(|\hat{r}-r_0|) \geq \left[\frac{E\left[\left(\frac{\partial F(r)}{\partial x}\right)^2 + \left(\frac{\partial F(r)}{\partial y}\right)^2\right]}{E\left[\left(\frac{\partial F(r)}{\partial x}\right)^2\right] E\left[\left(\frac{\partial F(r)}{\partial y}\right)^2\right] - \left\{E\left[\frac{\partial F(r)}{\partial x}\,\frac{\partial F(r)}{\partial y}\right]\right\}^2} \right]_{r=r_0} \tag{5}$$

where

$$F(r) = \sum_{ij} k_{ij} \log \lambda_{ij}(r)$$

and $E[x]$ denotes the expectation value (over $\{k_{ij}\}$) of the variable x. By using the assumption that k_{ij} are independent and Poisson-distributed, the lower bound can be further reduced to

$$\mathrm{Var}(|\hat{r}-r_0|) \geq \left[\frac{\sum_{ij} \frac{(\partial\lambda_{ij}(r)/\partial y)^2 + (\partial\lambda_{ij}(r)/\partial x)^2}{\lambda_{ij}(r)}}{\left[\sum_{ij} \frac{(\partial\lambda_{ij}(r)/\partial y)^2}{\lambda_{ij}(r)}\right]\left[\sum_{ij} \frac{(\partial\lambda_{ij}(r)/\partial x)^2}{\lambda_{ij}(r)}\right] - \left[\sum_{ij} \frac{(\partial\lambda_{ij}(r)/\partial x)(\partial\lambda_{ij}(r)/\partial y)}{\lambda_{ij}(r)}\right]^2} \right]_{r=r_0} \tag{6}$$

Given the source intensity distribution, $I_0(\rho)$, the variance of tracking error can be calculated. It should be noted that the CRLB is a lower bound on the estimator error. However, it provides an analytically tractable expression, and is therefore very useful in estimating the performance of the tracking system.

Note that the variance in Eq. (6) is in general a function of r_0, the actual geometric center. Furthermore, the variance depends on both the shape of the image as well as its intensity. To study the effect of increasing source intensity, or equivalently, increasing the integration time, one can normalize the receiver count parameter $\{\lambda_{ij}(r)\}$ as

$$\lambda_{ij}(r) = N_0 g_{ij}(r) \tag{7}$$

where N_0 is the average number of photons received over the tracking sensor, and $g_{ij}(r)$ is the fraction of photons that falls onto the $(i,j)^{th}$ pixel. By definition,

$$\sum_{ij} g_{ij}(r) = 1$$

Given the above definition, the CRLB can be written as

$$\mathrm{Var}(|\hat{r}-r_0|) \geq \frac{1}{N_0} \left[\frac{\sum_{ij} \frac{(\partial g_{ij}(r)/\partial y)^2 + (\partial g_{ij}(r)/\partial x)^2}{g_{ij}(r)}}{\left[\sum_{ij} \frac{(\partial g_{ij}(r)/\partial y)^2}{g_{ij}(r)}\right]\left[\sum_{ij} \frac{(\partial g_{ij}(r)/\partial x)^2}{g_{ij}(r)}\right] - \left[\sum_{ij} \frac{(\partial g_{ij}(r)/\partial x)(\partial g_{ij}(r)/\partial y)}{g_{ij}(r)}\right]^2} \right]_{r=r_0}$$

$$= \frac{1}{N_0} G(r_0) \tag{8}$$

It is easily seen that the performance of the tracking system improves with increasing signal power, N_0. The function $G(r_0)$ in Eq. (8) depends only on the shape of the image, and not its intensity. Figure 1 shows the values of $G(r_0)$ for two very simple cases. For a more general image shape, $G(r_0)$ is very difficult to calculate. However, it can be seen from Eq. (8) that the lower bound in error variance is minimized when

$$\left[\sum_{ij}\frac{(\partial g_{ij}(r)/\partial x)^2}{g_{ij}(r)}\right]_{r=r_0} \gg 1 \tag{9a}$$

$$\left[\sum_{ij}\frac{(\partial g_{ij}(r)/\partial y)^2}{g_{ij}(r)}\right]_{r=r_0} \gg 1 \tag{9b}$$

and

$$\left\{\left[\sum_{ij}\frac{(\partial g_{ij}(r)/\partial x)(\partial g_{ij}(r)/\partial y)}{g_{ij}(r)}\right]_{r=r_0}\right\}^2 \ll \left[\sum_{ij}\frac{(\partial g_{ij}(r)/\partial x)^2}{g_{ij}(r)}\right]_{r=r_0}\left[\sum_{ij}\frac{(\partial g_{ij}(r)/\partial y)^2}{g_{ij}(r)}\right]_{r=r_0} \tag{9c}$$

The expressions in Eqs. (9a) and (9b) are greater for images with high contrasts, i.e., images that contain pixels with high $|\partial g_{ij}(r)/\partial x|$ and $|\partial g_{ij}(r)/\partial y|$. Since the reference image is usually much brighter than the background, the partial derivatives are greater near the image border. It follows that the CRLB is smaller for images with better defined borders, i.e., images for which $\partial g_{ij}/\partial x$ and $\partial g_{ij}/\partial y$ are large.

III. Albedo Variation

The derivation above shows that, when the source intensity distribution is known, the variance in estimating the image location decreases linearly with increasing signal power. For a sufficiently bright source, the variance is negligible.

Unfortunately, the derivation that leads to Eq. (8) assumes that the exact source intensity distribution $I_0(r)$ is known at the receiver. In some cases, the intensity distribution of the receiver can be quite unpredictable. For example, for a deep-space vehicle using the sun-lit Earth as the pointing reference, the albedo variation of the Earth can be caused by weather patterns and changing surface conditions. Furthermore, these conditions are in general time-varying so that they cannot be expected to remain constant.

In order to quantify the effect of intensity uncertainty on the spatial tracking subsystem, some assumptions on the intensity error distribution are required. For this analysis, it is assumed that the estimated source intensity distribution $\hat{I}(\rho)$ differs from the actual source intensity $I_0(\rho)$ by a small amount $I_1(\rho)$. Furthermore, it will be assumed that $I_1(\rho)$ is a zero-mean Gaussian random process with an autocorrelation function

$$\phi(\rho,\rho') \equiv \langle I_1(\rho) I_1(\rho')\rangle \tag{10}$$

The maximum likelihood estimator must estimate the location of the image based on an incomplete estimate of the source distribution. In other words, the estimate $\hat{r}$ for the image center can be written as

$$\hat{r} = \arg\left[\max_r P(\{k_{ij}\}\,|\,r,\hat{I}(\rho))\right] = \arg\left\{\max_r F'(r)\right\} \tag{11}$$

where $F'(r) = \log[P(\{k_{ij}\}|r,\hat{I}(\rho))]$ is the likelihood function, and we have used the fact that a logarithmic transformation does not affect the location of a functional maximum.

Given the formulation of the estimator in Eq. (11), the variance of the estimation error can again be given in terms of its Cramer-Rao lower bound as in [1]

$$\mathrm{Var}(|\hat{r}-r_0|) \geq \left[\frac{E\left[\left(\frac{\partial F'(r)}{\partial x}\right)^2 + \left(\frac{\partial F'(r)}{\partial y}\right)^2\right]}{E\left[\left(\frac{\partial F'(r)}{\partial x}\right)^2\right] E\left[\left(\frac{\partial F'(r)}{\partial y}\right)^2\right] - \left\{E\left[\frac{\partial F'(r)}{\partial x}\,\frac{\partial F'(r)}{\partial y}\right]\right\}^2}\right]_{r=r_0} \tag{12}$$

The expectation in Eq. (12) is, in general, very complicated since the detector photocounts $\{k_{ij}\}$ are conditional Poisson-distributed random variables. Given the estimated intensity $\hat{I}(\rho)$, the actual source intensity distribution $I_0(\rho)$ can be modeled as a random variable with mean $\hat{I}(\rho)$. That is, the mean photocount expected over the $(i,j)^{th}$ pixel, λ_{ij}, is a random variable with mean

$$\hat{\lambda}_{ij}(r) = \int_{I_{ij}} \hat{I}(\rho - r)\, d^2 r \tag{13}$$

When the source intensity is sufficiently high, or over a sufficiently long integration time, the fluctuation in the Poisson count statistics will be small compared to its mean. In this case one can approximate the detector photocount in Eq. (11) by its mean value, and the likelihood function in Eq. (11) can be reduced to

$$F'(r) = \log\,[P(\{K_{ij}\}\,|\,r, \hat{I}(\rho))] \approx \ln[P(\{\lambda_{ij}\}\,|\,r, \{\hat{\lambda}_{ij}(r)\})] \tag{14}$$

By using this approximation, the probability given in Eq. (11) can be interpreted as the probability of receiving $\lambda_{ij}(r_0)$ given the source intensity distribution $\{\hat{\lambda}_{ij}(\rho)\}$. Since it is assumed that $I_0(\rho)$ differs from $\hat{I}(\rho)$ by a zero-mean Gaussian process $I_1(\rho)$, it follows that the probability in Eq. (14) can be written as

$$P(\{\lambda_{ij}\}\,|\,r, \hat{I}(\rho)) = \frac{1}{\sqrt{(2\pi)^M |\sigma_{ij,\ell m}|}} \cdot \exp\left[-\frac{1}{2}\sum_{ij,\ell m}(\lambda_{ij} - \hat{\lambda}_{ij}(r))\,\sigma^{-1}_{ij,\ell m}(\lambda_{\ell m} - \hat{\lambda}_{\ell m}(r))\right] \tag{15}$$

where M is the total number of pixels used in the decision, and $\sigma^{-1}_{ij,\ell m}$ and $|\sigma_{ij,\ell m}|$ denote the matrix inverse and the determinant of the covariance matrix $\sigma_{ij,\ell m}$, respectively. The elements of the covariance matrix can be given by

$$\sigma_{ij,\ell m}(r) = \langle (\hat{\lambda}_{ij}(r) - \lambda_{ij}(r))(\hat{\lambda}_{\ell m}(r) - \lambda_{\ell m}(r)) \rangle = \int_{A_{ij}} \int_{A_{\ell m}} \phi(\rho - r, \rho' - r)\, d^2\rho\, d^2\rho' \tag{16}$$

By differentiating the likelihood function in Eq. (14) and taking the expectation values, the Cramer-Rao bound of the estimator error variance can be given in terms of Eq. (12) (see Appendix) where

$$E\left[\left(\frac{\partial F'(r)}{\partial x}\right)^2\right] = \left[\sum_{ij,\ell m}\left(\left(\frac{\partial \hat{\lambda}_{ij}}{\partial x}\right)\left(\frac{\partial \hat{\lambda}_{\ell m}}{\partial x}\right)[\sigma^{-1}]_{ij,\ell m} + \left(\frac{\partial \sigma_{ij,\ell m}}{\partial x}\right)\left(\frac{\partial [\sigma^{-1}]_{ij,\ell m}}{\partial x}\right)\right)\right] \tag{17a}$$

$$E\left[\left(\frac{\partial F'(r)}{\partial y}\right)^2\right] = \left[\sum_{ij,\ell m}\left(\left(\frac{\partial \hat{\lambda}_{ij}}{\partial y}\right)\left(\frac{\partial \hat{\lambda}_{\ell m}}{\partial y}\right)[\sigma^{-1}]_{ij,\ell m} + \left(\frac{\partial \sigma_{ij,\ell m}}{\partial y}\right)\left(\frac{\partial [\sigma^{-1}]_{ij,\ell m}}{\partial y}\right)\right)\right] \tag{17b}$$

$$E\left[\left(\frac{\partial F'(r)}{\partial x}\right)\left(\frac{\partial F'(r)}{\partial y}\right)\right] = \left[\sum_{ij,\ell m}\left(\left(\frac{\partial \hat{\lambda}_{ij}}{\partial x}\right)\left(\frac{\partial \hat{\lambda}_{\ell m}}{\partial y}\right)[\sigma^{-1}]_{ij,\ell m} + \left(\frac{\partial \sigma_{ij,\ell m}}{\partial x}\right)\left(\frac{\partial [\sigma^{-1}]_{ij,\ell m}}{\partial y}\right)\right)\right] \tag{17c}$$

Equations (12) and (17) together present an analytical form of the mean square estimator error. Given the estimated intensity pattern $\hat{I}(\rho)$, and the intensity correlation function $\phi(\rho, \rho')$, the lower bound for the variance in estimating the image location can be calculated. Unfortunately, for general distributions of $\hat{I}(\rho)$ and $\phi(\rho, \rho')$, the expressions in Eqs. (17a-c) are very difficult to evaluate. In order to obtain some insight into the functional dependence of tracking error variance and the source intensity error, some simplifications are required. In the following we shall present several simple cases that will illustrate these dependencies.

Example 1: White intensity noise with spectral density γ^2. That is,

$$\langle I_1(\rho)\, I_1(\rho') \rangle = \frac{\gamma^2}{A}\,\delta(\rho - \rho') \tag{18}$$

where A is the area of a pixel element, and $\delta(x)$ is the Dirac delta function. An example of this type of intensity uncertainty is the random dark counts from the tracking detector. By using Eq. (16), the correlation matrix can be calculated. The result is

$$\sigma_{ij,\ell m} = \gamma^2 \delta_{i\ell}\, \delta_{jm} \tag{19}$$

where δ_{ij} is the Kronecker delta. By substituting Eq. (20) into Eqs. (17 a-c), and using the fact that the $\sigma_{ij,\ell m}$ variables do not depend on r, the CRLB can be reduced to

$$\mathrm{Var}(|\hat{r} - r_0|) \geq$$

$$\frac{\gamma^2}{N_0} \frac{\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial x}\right)^2 + \sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial y}\right)^2}{\left[\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial x}\right)^2\right]\left[\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial y}\right)^2\right] - \left[\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial x}\right)\left(\frac{\partial \hat{g}_{ij}}{\partial y}\right)\right]^2} \tag{20}$$

where the total detector SNR, N_0, is factored out by making the substitution $\lambda_{ij}(r) = N_0\, g_{ij}(r)$. It is easily seen from Eq. (20) that the variance in estimating the image location is directly proportional to the uncertainties in source intensity distribution γ^2. Furthermore, the variance of the estimator error decreases with increasing N_0 and, at a very high signal count, the variance is negligible.

Example 2: White intensity fluctuation with spectral density that is proportional to the total signal intensity. In other words,

$$\phi(\rho, \rho') = \frac{\gamma^2 N_0^2}{A} \delta(\rho - \rho') \tag{21}$$

In this case the uncertainty in image brightness is proportional to the intensity of the image. An example of this type of intensity uncertainty is the unknown albedo variation of the source. An increase in the integration time at the tracking detector will only result in a proportional increase of the uncertainty. Under this condition, the CRLB reduces to

$$\mathrm{Var}(|\hat{r} - r_0|) \geq$$

$$\gamma^2 \left[\frac{\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial x}\right)^2 + \sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial y}\right)^2}{\left[\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial x}\right)^2\right]\left[\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial y}\right)^2\right] - \left[\sum_{ij}\left(\frac{\partial \hat{g}_{ij}}{\partial x}\right)\left(\frac{\partial \hat{g}_{ij}}{\partial y}\right)\right]^2}\right]_{r=r_0} \tag{22}$$

Note that even though the CRLB still depends linearly on γ^2, the lower bound no longer depends on the total detector SNR. Consequently, Eq. (22) represents an irreducible error floor for the estimator. For the simple test patterns shown in Fig. 1, the magnitude of the CRLB can be easily calculated to be $\gamma^2(b + 2a)^3/a$ and $\gamma^2 \pi^2 a^2/2$, respectively.

IV. Tracking System Using Sun-Lit Earth as a Pointing Reference

In a typical spatial tracking subsystem, the transmitter pointing information is derived from the image location of the reference source. The reference source can be either an uplink beacon laser, or the sun-lit Earth. The reference signal is received by the telescope and, after spatial and frequency filtering to cut down the background noise, is focused onto the tracking detector. The tracking detector is usually a focal-plane array which spans the field-of-view (FOV) of the receiving optics, and can be implemented using a large-format CCD.

Deriving the angular coordinate of the reference source is equivalent to determining the position of its image. Since objects with angular separation less than the resolution limits cannot be resolved by the receiving optics, it is generally desirable to design the optics such that the pixel size on the focal plane array corresponds to the resolution limit of the telescope. Such a design provides maximum spatial information with a minimum number of pixels. For a spacecraft at a distance of 2.5 AU using a 10-cm transmitter, the image of the Earth will span roughly 4–5 pixels. The required pointing accuracy, on the other hand, is less than 1/10 of the transmitted beamwidth. Since the angular resolution of the telescope is equivalent to the minimal divergence of the transmitted optical signal, deriving the desired pointing accuracy of 1/10 the beamwidth is therefore equivalent to locating the position of the receiving station to within 1/10 of a pixel size based on the Earth image on the focal plane array.

When the intensity distribution of the pointing reference is known, as would be the case when an uplink beacon is used, the performance of the receiver is given by Eq. (6). The expected detector SNR, N_0, can be calculated using simple link analysis. By using the additional assumption that atmospheric scattering limits the angular divergence of the uplink beacon to about 20 μrad, the detector SNR can be approximated by

$$N_0 \approx 5 \times 10^2 P_s D_R^2 T/z^2 \tag{23}$$

where P_s is the beacon power in watts, D_R is the receiver diameter in meters, T is the integration time, and z is the link distance in AUs. Equation (23) was derived by assuming a 532-nm uplink beacon, and that the losses in optics and detector are

negligible. The actual signal count can be much lower than that given by Eq. (23) due to these losses.

It is easily seen from Eq. (23) that, for a tracking detector operating at 1 KHz using a 10-cm diameter receiver at 1 AU, the required signal power for a 20-dB SNR is 20 KW! Obviously, such a high power can be very costly to achieve. And if higher SNR is desired, the required beacon power can be even higher.

An alternative is to use the sun-lit Earth as a pointing reference. Sunlight reflection off the Earth can provide a large amount of signal power at the tracking detector. In fact, a simple calculation shows that a detector SNR higher than 10^5 can be easily achieved. As a result, tracking error due to Poisson-count statistics is negligible. Unfortunately, the albedo of the Earth cannot be specified precisely. Cloud-cover can alter the surface reflectivity significantly, and a snow-covered surface can reflect up to one order of magnitude more sunlight than an exposed terrain. To further complicate the problem, these conditions are changing in time so that it is almost impossible to derive an accurate estimate of the Earth's albedo. Since the uncertainty is inherent to the brightness of the source, increasing the integration will only result in corresponding increases in that uncertainty. Consequently, there will be an irreducible error floor on the tracking system performance.

The actual impact of albedo uncertainty on the tracking system performance depends, of course, on the magnitude of the uncertainty and its spatial correlation. In general, the CRLB given by Eqs. (12) and (17) is very difficult to compute. For the simple test patterns shown in Fig. 1, a minimal standard deviation of 0.1 pixel requires that γ be less than $0.1\sqrt{a/(b+2a)^3}$ and $0.045/a$, respectively. If these examples are representative, then we will need to know the intensity to within 10 percent of the true value in order to limit the pointing error to within 0.1 pixel. Since the average Earth albedo variation is much more than 10 percent, a simple ML estimator cannot be expected to achieve the desired tracking accuracy.

V. Conclusions

Because of the large link distance involved, it is desirable that the optical communication package aboard a planetary spacecraft derive its pointing reference directly from the image of the sun-lit Earth on the tracking detector. Given the detector photocounts and the prior knowledge of the image intensity distribution, the maximum-likelihood spatial acquisition algorithm can be derived. The performance of the maximum-likelihood algorithm was analyzed by calculating the Cramer-Rao bound on the variance of acquisition error. It is shown that, when the intensity distribution of the pointing reference can be precisely characterized, the variance in estimating the receiver angular location decreases with increasing image intensity or detector exposure time. On the other hand, when the intensity distribution is not known in sufficient detail, the spatial tracking error variance will not decrease indefinitely with increasing exposure time. For a planetary spacecraft using the Sun-lit Earth as a pointing reference, the Earth's albedo cannot be precisely specified because of changing weather and ground conditions. Consequently, the ML algorithm which derives the pointing reference based on the detector photocounts cannot be expected to provide an accurate pointing reference.

It should be noted, however, that the results presented in this study were derived by assuming that the receiver derives its pointing reference based on a single frame of the image. In other words, the receiver estimates the image location, r, based on the receiver count statistics and an assumed source intensity distribution, $\hat{I}(\rho)$. When multiple images of the pointing reference are available, the receiver can then jointly estimate the true source distribution, $I_0(\rho)$, and the image location, r. For such a receiver, the tracking system performance will not be limited by the intensity uncertainty. In fact, with a sufficiently large number of look angles, one could derive an estimate $\hat{I}(\rho)$ which closely approximates $I_0(\rho)$ and, consequently, the variance in estimating r could be reduced to an acceptable level.

References

[1] K. Winick, "Cramer-Rao Lower Bound on the Performance of Charge-Coupled-Device Optical Position Estimators," *J. Opt. Soc. Am. A*, vol. 3, no. 11, pp. 1809–1815, November 1986.

[2] M. Win, "Acquisition and Tracking for Optical Communications," submitted to OE-LASE'89, Los Angeles, California, January 1989.

[3] H. Van Tree, *Detection, Estimation, and Modulation Theory*, New York: Wiley, 1968.

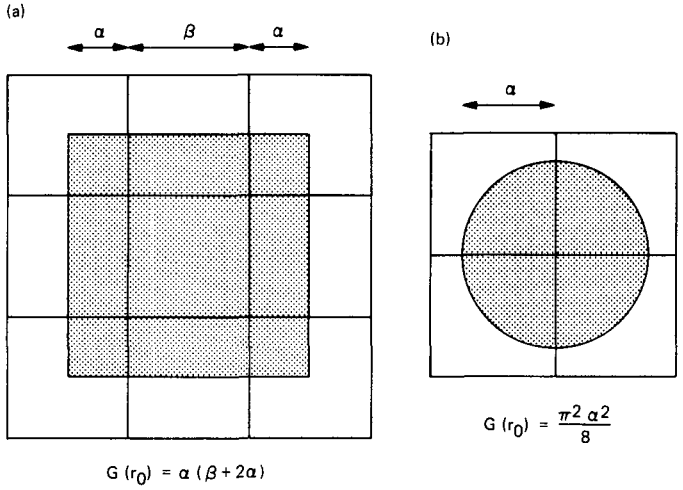


Fig. 1. The value of $G(r_0)$ for two simple examples of source intensity distributions over the detector focal plane. The solid grids represent the CCD pixels, and the shaded areas represent the image of the pointing reference. The image intensities are assumed to be uniformly distributed.

Appendix
Derivation of Equation (16)

The logarithm of the joint probability distribution $P(\{\lambda_{ij}\}|\, r, I(\rho))$, is

$$\ln\left[P(\{\lambda_{ij}\}|\, r, \hat{I}(\rho))\right] = -\frac{1}{2}\ln|\sigma_{ij,\ell m}| - \frac{1}{2}\sum_{ij,\ell m}(\lambda_{ij} - \hat{\lambda}_{ij}(r))[\sigma^{-1}]_{ij,\ell m}(\lambda_{\ell m} - \hat{\lambda}_{\ell m}) \tag{A-1}$$

By expanding $E\,[(\partial \ln(P(\{\lambda_{ij}\}|\, r, I(\rho))/\partial x)^2]$ into an integral form, and using the fact that

$$\int(\partial^2 P(\{\lambda_{ij}\})/\partial x^2)\, d\,\{\lambda_{ij}\} = 0$$

it follows that

$$E\left[\left(\frac{\partial \ln\,[P(\{\lambda_{ij}\}|\, r, I(\rho))]}{\partial x}\right)^2\right] = \int\left(\frac{\partial P(\{\lambda_{ij}\})/\partial x}{P(\{\lambda_{ij}\})}\right)^2 P(\{\lambda_{ij}\})\, d\{\lambda_{ij}\}$$

$$= \int\left\{\left(\frac{\partial P(\{\lambda_{ij}\})/\partial x}{P(\{\lambda_{ij}\})}\right)^2 + \left(\frac{\partial^2 P(\{\lambda_{ij}\})/\partial x^2}{P(\{\lambda_{ij}\})}\right)\right\} P(\{\lambda_{ij}\})\, d\{\lambda_{ij}\}$$

$$= \int\frac{\partial^2 \ln\,[P(\{\lambda_{ij}\}|r, I(\rho))]}{\partial x^2} P(\{\lambda_{ij}\})\, d\{\lambda_{ij}\}$$

$$= -E\left[\frac{\partial^2 \ln\,[P(\{\lambda_{ij}\}|r, I(\rho))]}{\partial x^2}\right] \tag{A-2}$$

By differentiating Eq. (A-1) and using the fact that $E\,[\partial P(\lambda_{ij}|\, r, \hat{I}(\rho))/\partial x]$ is equal to zero for all r, one can show that

$$\frac{\partial}{\partial x} E\left[\frac{\partial}{\partial x} P(\{\lambda_{ij}\}|\, r, \hat{I}(\rho))\right] = -\frac{\partial^2}{\partial x} \ln|\sigma_{ij,\ell m}| - \frac{1}{2}\sum\sigma_{ij,\ell m}\frac{\partial^2[\sigma^{-1}]_{ij,\ell m}}{\partial x^2} - \frac{1}{2}\sum\sigma_{ij,\ell m}\frac{\partial \sigma_{ij,\ell m}}{\partial x}\frac{\partial[\sigma^{-1}]_{ij,\ell m}}{\partial x}$$

$$\equiv 0 \tag{A-3}$$

In deriving Eq. (A-3), we have used the assumption that $\{\lambda_{ij}\}$ are Gaussian distributed with mean $\{\hat{\lambda}_{ij}\}$.

By differentiating Eq. (A-1) and taking the expectation values of λ_{ij}, Eq. (A-2) is reduced to

$$E\left[\left(\frac{\partial}{\partial x}\ln[P(\{\lambda_{ij}\}|\, r, \hat{I}(\rho))]\right)^2\right] = \frac{1}{2}\frac{\partial^2}{\partial x^2}\ln|\sigma_{ij,\ell m}| + \frac{1}{2}\sum_{ij,\ell m}\sigma_{ij,\ell m}\frac{\partial^2[\sigma^{-1}]_{ij,\ell m}}{\partial x^2} + \frac{1}{2}\sum_{ij,\ell m}\left(\frac{\partial\hat{\lambda}_{ij}}{\partial x}\right)[\sigma^{-1}]_{ij,\ell m}\left(\frac{\partial\hat{\lambda}_{\ell m}}{\partial x}\right) \tag{A-4}$$

By substituting Eq. (A-3) into (A-4), it is seen that

$$E\left[\left(\frac{\partial}{\partial x}\ln[P(\{\lambda_{ij}(r_0)\}\mid r,\hat{I}(\rho))]\right)^2\right] = \left[\sum_{ij,\ell m}\left(\left(\frac{\partial\lambda_{ij}}{\partial x}\right)\left(\frac{\partial\lambda_{\ell m}}{\partial x}\right)\,[\sigma^{-1}]_{ij,\ell m} + \left(\frac{\partial\sigma_{ij,\ell m}}{\partial x}\right)\left(\frac{\partial[\sigma^{-1}]_{ij,\ell m}}{\partial x}\right)\right)\right] \tag{A-5}$$

Similarly, it can be shown that

$$E\left[\left(\frac{\partial}{\partial y}\ln[P(\{\lambda_{ij}(r_0)\}\mid r,\hat{I}(\rho))]\right)^2\right] = \left[\sum_{ij,\ell m}\left(\left(\frac{\partial\lambda_{ij}}{\partial y}\right)\left(\frac{\partial\lambda_{\ell m}}{\partial y}\right)\,[\sigma^{-1}]_{ij,\ell m} + \left(\frac{\partial\sigma_{ij,\ell m}}{\partial y}\right)\left(\frac{\partial[\sigma^{-1}]_{ij,\ell m}}{\partial y}\right)\right)\right] \tag{A-6}$$

and

$$E\left[\left(\frac{\partial\ln[P(\{\hat{\lambda}_{ij}(r_0)\}\mid r,\hat{I}(\rho))]}{\partial x}\right)\left(\frac{\partial\ln[P(\{\hat{\lambda}_{ij}(r_0)\}\mid r,\hat{I}(\rho))]}{\partial y}\right)\right] = \left[\sum_{ij,\ell m}\left(\left(\frac{\partial\lambda_{ij}}{\partial x}\right)\left(\frac{\partial\lambda_{\ell m}}{\partial y}\right)\,[\sigma^{-1}]_{ij,\ell m} + \left(\frac{\partial\sigma_{ij,\ell m}}{\partial x}\right)\left(\frac{\partial[\sigma^{-1}]_{ij,\ell m}}{\partial y}\right)\right)\right] \tag{A-7}$$

A Preliminary Weather Model for Optical Communications Through the Atmosphere

K. S. Shaik
Communications Systems Research Section

A preliminary weather model is presented for optical propagation through the atmosphere. It can be used to compute the attenuation loss due to the atmosphere for desired link availability statistics. The quantitative results that can be obtained from this model provide good estimates for the atmospheric link budget necessary for the design of an optical communication system. The result is extended to provide for the computation of joint attenuation probability for n *sites with uncorrelated weather patterns.*

I. Introduction

During the past few decades, the pioneering work of Tatarski [1], Fried [2,3], Hufnagel and Stanley [4], and Ishimaru [5], has set the stage for the development of a consistent theory to evaluate the performance of optical communications and imaging through the atmosphere. However, there is a conspicuous lack of weather models that may be used to obtain even a first-order estimate of statistics for the availability of an optical communications link under ambient weather conditions.

The light energy of a beam dissipates as it travels through the atmosphere, due to scattering and absorption. If it can be assumed that (1) attenuation loss is independent of radiation intensity and (2) the absorbing and scattering events occur independently of each other, the atmospheric attenuation can be expressed by the Bouguer-Lambert law:

$$I(\nu) = I_0(\nu) \exp\left[-\int_0^Z \gamma(\nu, z)\, dz\right] \tag{1}$$

where $I(\nu)$ is the observed irradiance at optical frequency ν, $I_0(\nu)$ is the irradiance that would have been observed if the beam were propagating in a vacuum at a distance Z from the source, and $\gamma(\nu, z)$ is the total extinction coefficient due to scattering and absorption by the atmospheric constituents at position z. The magnitude of the argument of the exponential in the above equation is known as the optical depth or optical thickness, τ, of the medium, i.e.,

$$\tau = \int_0^Z \gamma(\nu, z)\, dz \tag{2}$$

Experiments [6] with artificial fog and smoke and diluted solutions of milk show that the Bouguer-Lambert law holds well for optical thickness $\tau \leqslant 12$. The optical depth can be simply related to the attenuation loss, L, in dB by the following approximation:

$$L \approx 4.34\tau \tag{3}$$

The distribution of gas molecules, clouds, fog, haze, aerosols, and other particulates in the transmission path influence the computation of attenuation loss. These phenomena at best can be interpreted on a statistical basis. The system designer may have to be content with a probability for which the attenuation of the atmosphere due to ambient weather is less than some critical value L dB. Hence we define the probability, $w_1(L)$, of the random weather variable, W, for a single site as

$$w_1(L) = P_W(l \leq L) \tag{4}$$

where l is the loss variable. $w_1(L)$ is the fraction of weather conditions at any given site for which the attenuated direct beam irradiance $I(\nu) \geq I_0(\nu) \exp[-0.23\, L]$. The weather model developed below can be used to compute this probability. The result is also extended to obtain the joint weather probability, $w_n(L)$, that at least one of the sites has extinction loss $l \leq L$ for n independent sites receiving simultaneously.

II. Weather Model

The following observations are made in order to arrive at the preliminary weather model:

(1) It is assumed that for some fraction of the time p, where $0 \leq p \leq 1$, clear weather conditions hold. This number can be determined approximately from existing data on cloud cover and visibility for potential sites. An analysis of two years of GOES satellite data by Wylie and Menzel [7] shows that the probability of having clear weather in the southwestern U.S. is over 60% (Fig. 1).

(2) The attenuation under clear weather conditions is due to scattering and absorption by air molecules and sparse particulate matter in the upper atmosphere. The attenuation due to molecular absorption is approximately 0.5 dB [6]. The attenuation due to molecular scattering is of the same order. Given that the atmosphere is never totally free of aerosols and thin cirrus clouds, on an average clear day the attenuation loss would be in the range of 1 to 3 dB. For simplicity, let us assume that the minimum attenuation loss due to the atmosphere is 3 dB. Hence, the model defines clear-air atmosphere, which occurs with probability p, as having an attenuation loss $L_0 = 3$ dB. Note that this attenuation loss refers to near-zenith propagation through the entire atmosphere. For any zenith angle θ, the path attenuation can be written as $L_0 \sec\theta$. However, the remainder of the discussion assumes near-zenith propagation paths, i.e., $\theta = 0$.

(3) The probability of not having clear weather is $q = 1 - p$. The attenuation loss of the atmosphere increases rapidly with increasing concentration of aerosols, fog, haze, and clouds. Optical attenuations $L > 1000$ dB are possible, where the validity of the Bouguer-Lambert law is questionable. Table 1 shows typical ranges of attenuation values for various types of clouds [8]. Since it is unlikely that an optical communication system will be designed for $L > 30$ dB, the Bouguer-Lambert law serves as a good approximation. As shown in Fig. 2, the probability of opaque cloud cover over the southwestern U.S. is 20% [7]. Let us take a pessimistic view and assume that attenuation loss due to all opaque clouds is 100 dB or higher, and use this assumption to estimate one of the model parameters.

(4) It is further assumed that the attenuation of the beam has an exponential distribution for $L \geq L_0$. The hypothesis is supported by visual observations, and also by the experimental data collected at Goldstone in California for atmospheric propagation at 8 to 10 Ghz.[1]

With the foregoing assumptions in mind, it is now possible to postulate a weather model. The weather cumulative distribution function (CDF) for a single site can be modeled as

$$w_1(L) = 1 - q \exp[-0.23\, b(L - L_0)] \qquad (L \geq L_0) \tag{5}$$

where $w_1(L)$ is defined in Eq. (3) above, q is the probability fraction when the weather is not clear, b is a site-dependent parameter to model the slope of the CDF curve, and L_0 is the minimum attenuation loss due to the atmosphere. Note that $w_1(L = L_0) = p$. The value of b may vary with geographical location and altitude, and can be inferred from observed visibility and extinction loss data for potential receiving sites. For the southwestern U.S. region, $q = 0.4$, and from the third item above, the attenuation loss is 100 dB when $w_1(L = 100) = 0.8$. Using these values in Eq. (5), we find that $b = 0.03$. This estimate of the parameter b will be used in numerical examples later.

Equation (5) can be recast in a more familiar form for an optical communication system designer to give the attenuation loss in terms of the weather probability $w_1(L)$, i.e.,

$$L = \begin{cases} L_0, & \text{for } w_1(L) \leq p \\ L_0 + \dfrac{1}{0.23\, b} \ln\left[\dfrac{q}{1 - w_1(L)}\right], & \text{for } w_1(L) > p \end{cases} \tag{6}$$

[1] JPL Internal Document 810-5, Rev. D, 1988.

Figure 3 is a plot of attenuation loss as a function of the weather probability for the southwestern region of the U.S., for which $q = 0.4$, $b = 0.03$, and $L_0 = 3$ dB. These estimates, needless to say, are quite important for the system designer to determine the link budget for loss due to the atmosphere for Earth-space optical communication paths.

It is also possible to extend the result to n sites receiving simultaneously. From Eq. (4), it is easy to see that the complement of the probability $w_1(L)$ is given by

$$P_W(l > L) = w_1^c(L) = q \exp[-0.23\, b\, (L - L_0)] \qquad (L \geqslant L_0) \qquad (7)$$

The joint probability of n sites, $w_n^c(L)$, that the attenuation loss $l_i > L$ for all i can be written as

$$w_n^c(L) = P_{W_1, W_2, \ldots, W_n}(l_1 > L, l_2 > L, \ldots, l_n > L) \qquad (8)$$

where the subscripts $1, 2, \ldots, n$ label the receiving sites. When the weather conditions for all sites are independent and identically distributed (IID), we have

$$w_n^c(L) = P_{W_1}(l_1 > L)\, P_{W_2}(l_2 > L) \ldots P_{W_n}(l_n > L) = [P_W(l > L)]^n \qquad (9)$$

Using Eq. (9), the joint probability that at least one site has attenuation $l \leqslant L$ is found to be

$$w_n(L) = 1 - w_n^c(L) = 1 - [q \exp[-0.23\, b\, (L - L_0)]]^n \qquad (L \geqslant L_0) \qquad (10)$$

For a single site with $p = 0.6$, $L_0 = 3$ dB, and $b = 0.03$, the probability that the attenuation $L \leqslant 3$ dB is $w_1(L = 3) = 0.60$. If three such IID sites are chosen, we have $w_3(L = 3) = 0.94$. In other words, if a system is designed to absorb an extinction loss of 3 dB, a three-site receiving network will be functional 94% of the time. Figure 4 plots the fraction of the total period under ambient weather conditions when the attenuation is $\leqslant 3$ dB as a function of the number of sites. Table 2 gives the expected dB loss for a desired link availability percentage for up to four joint receiving sites. It is, however, not very clear how the independence of weather patterns at various sites can be insured. It is known that the scale size of weather patterns is on the order of a few hundred kilometers, and this measure may be used to find sites with uncorrelated weather. Joint observation of weather parameters for the probable sites will be necessary to make a more accurate determination.

III. Conclusion

The virtue of the weather model presented here lies in its simplicity. It may be applied with ease to obtain a first-order magnitude of probabilities $w_n(L)$ for an optical system that must operate in the atmosphere. The computation of these probabilities, for example, will provide a good statistical estimate of the attenuation loss to an optical communication system designer for link availability.

The model does not consider frequency dependence, since it has been studied thoroughly and well documented in LOWTRAN computer code developed at the Air Force Geophysics Laboratory [9]. The model also disregards seasonal variations, which can be incorporated later when adequate data bases have been developed.

Acknowledgment

The author wishes to thank James K. Lesh for his comments and discussions, and also Dr. Chien-Chu Chen for his helpful suggestions.

References

[1] V. I. Tatarskii, *Wave Propagation in a Turbulent Medium*, New York: Dover, 1967.

[2] D. L. Fried, "Propagation of a Spherical Wave in a Turbulent Medium," *J. Opt. Soc. Am.*, vol. 57, pp. 175-180, 1967.

[3] D. L. Fried, "Limiting Resolution Looking Down Through the Atmosphere," *J. Opt. Soc. Am.*, vol. 56, pp. 1380-1384, 1966.

[4] R. E. Hufnagel and N. R. Stanley, "Modulation Transfer Function With Image Transmission Through Turbulent Media," *J. Opt. Soc. Am.*, vol. 54, pp. 52-61, 1964.

[5] A. Ishimaru, *Wave Propagation and Scattering in Random Media*, New York: Academic Press, 1977.

[6] V. E. Zuev, *Laser Beams in the Atmosphere*, New York: Consultants Bureau, 1982.

[7] D. P. Wylie and W. P. Menzel, "Cloud Cover Statistics Using VAS," *SPIE's OE-LASE'88 Symposium on Innovative Science and Technology*, Los Angeles, CA, January 10-15, 1988.

[8] The Technical Cooperation Program, Volume V— Laser Communications Workshop held in Australia, U.S. Government Printing Office no. 559-065/20988, 1985.

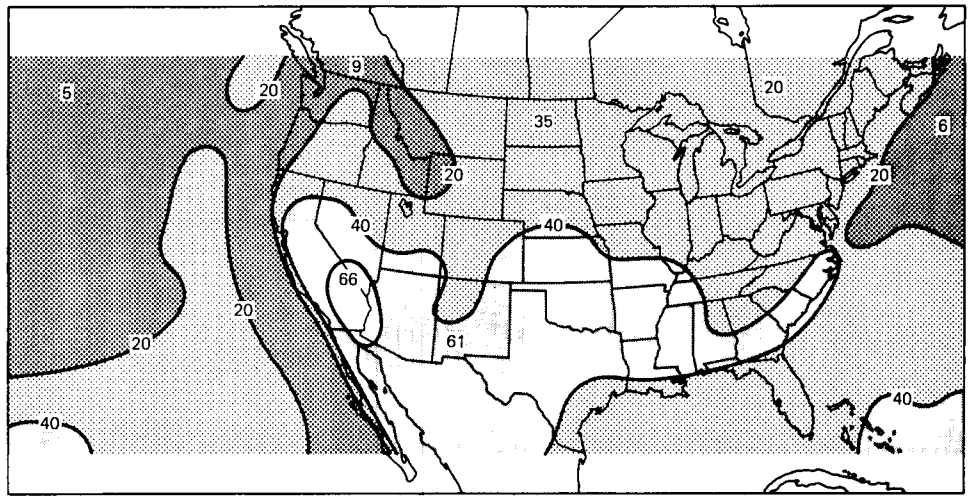
[9] F. X. Kneizys, et al., "Atmospheric Transmittance/Radiance: Computer Code LOWTRAN6," Rep. AFGL-TR-83-0187, 1983.

Table 1. Typical range of values for cloud attenuation (adapted from [8])

Cloud type	Cloud base height, km	Cloud thickness, km	Total vertical attenuation, dB
Stratus	0.1 – 0.7	0.2 – 0.8	26 – 350
Stratocumulus	0.6 – 1.5	0.2 – 0.8	3 – 100
Nimbostratus	0.1 – 1.0	2 – 3	260 – 1300
Altostratus/ cumulus	2 – 6	0.2 – 2	9 – 260
Cumulus	0.5 – 1.0	0.5 – 5	22 – 870
Cumulonimbus	0.5 – 1.0	2 – 12	260 – 5200
Cirriform (ice)	6 – 10	1.0 – 2.5	1 – 15
Fog	0	0 – 0.15	0 – 13

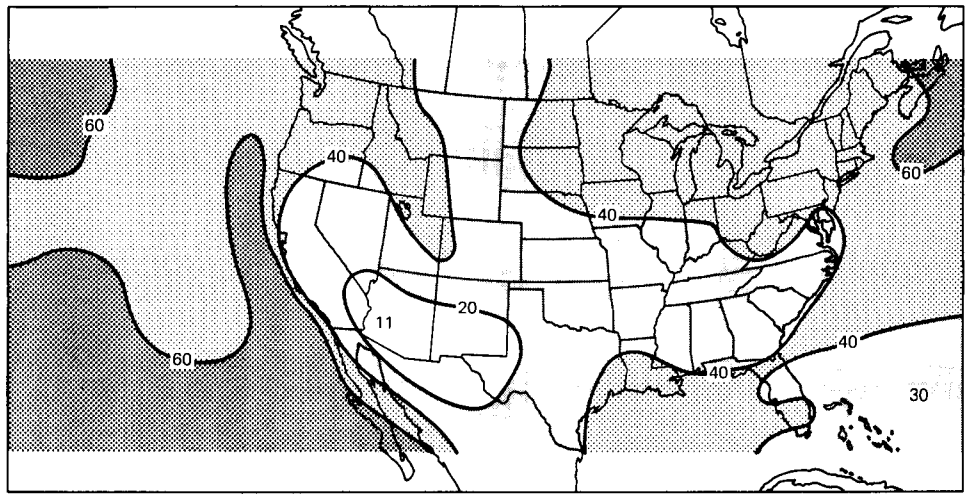
Table 2. Attenuation loss as a function of desired link availability for *n* sites receiving jointly

Percentage weather	Attenuation loss, L, dB			
	n = 1	n = 2	n = 3	n = 4
60.0	3.00	3.00	3.00	3.00
62.0	10.34	3.00	3.00	3.00
64.0	18.07	3.00	3.00	3.00
66.0	26.25	3.00	3.00	3.00
68.0	34.92	3.00	3.00	3.00
70.0	44.16	3.00	3.00	3.00
72.0	54.03	3.00	3.00	3.00
74.0	64.63	3.00	3.00	3.00
76.0	76.08	3.00	3.00	3.00
78.0	88.53	3.00	3.00	3.00
80.0	102.16	3.00	3.00	3.00
82.0	117.23	3.00	3.00	3.00
84.0	134.09	3.00	3.00	3.00
86.0	153.19	12.55	3.00	3.00
88.0	175.24	23.58	3.00	3.00
90.0	201.32	36.62	3.00	3.00
92.0	233.25	52.58	3.00	3.00
94.0	274.40	73.16	6.08	3.00
96.0	332.41	102.16	25.41	3.00
98.0	431.57	151.74	58.47	11.83
99.0	530.72	201.32	91.52	36.62



PROBABILITY OF CLEAR SKY, PERCENT

Fig. 1. Contour diagram obtained from 2 years of GOES satellite data showing the probability of clear sky over the United States [7].



PROBABILITY OF OPAQUE CLOUD COVER, PERCENT

Fig. 2. Contour diagram obtained from 2 years of GOES satellite data showing the probability of opaque cloud cover over the United States [7].

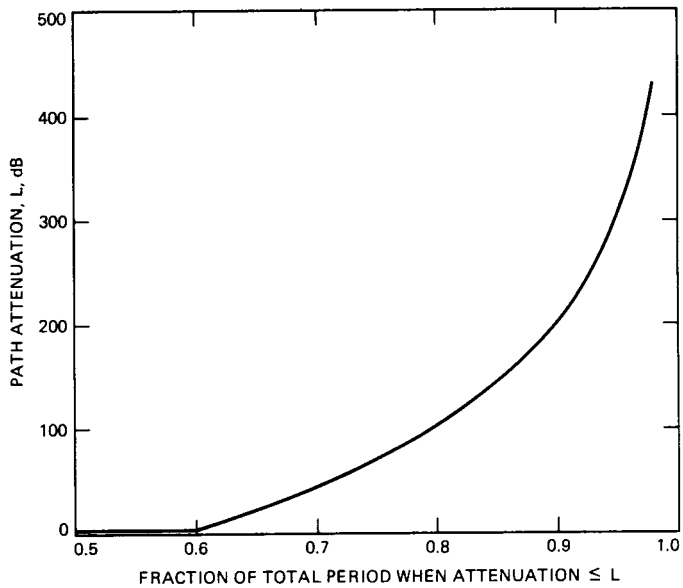


Fig. 3. Model curve for the path attenuation as a function of weather probability for a single site. Choice of parameters, L_0 = 3 dB, b = 0.03, and p = 0.6 represents estimates for the southwestern U.S.

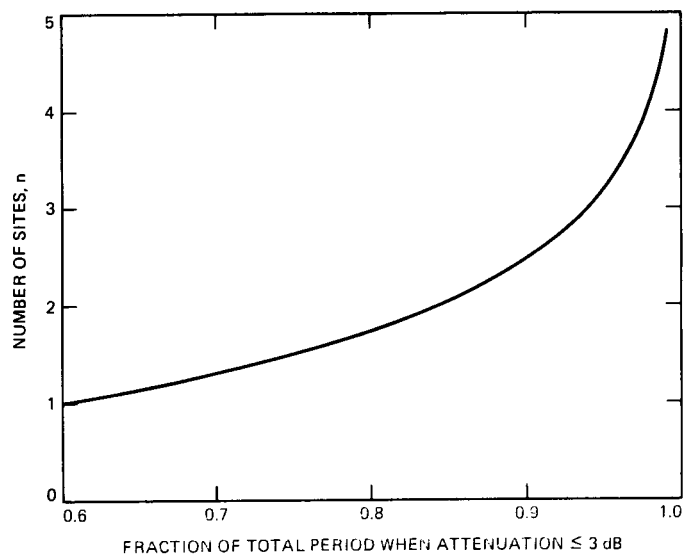


Fig. 4. Weather probability as a function of the number of joint receiving sites. L_0 = 3 dB, b = 0.03, p = 0.6, and L = 3 dB.

An Extended Kalman Filter Based Automatic Frequency Control Loop

S. Hinedi
Communications Systems Research Section

A novel Automatic Frequency Control (AFC) loop based on an Extended Kalman Filter (EKF) is introduced and analyzed in detail. The new scheme involves an EKF which operates on a modified set of data in order to track the frequency of the incoming signal. The algorithm can also be viewed as a modification to the well known cross-product AFC loop. A low carrier-to-noise ratio (CNR), high-dynamic environment is used to test the algorithm and the probability of loss-of-lock is assessed via computer simulations. The scheme is best suited for scenarios in which the frequency error variance can be compromised to achieve a very low operating CNR threshold. This technique can easily be incorporated in the Advanced Receiver (ARX), requiring minimum software modifications.

I. Introduction

The algorithm introduced and analyzed herein is another application of the theory of Kalman filters [1] to the problem of estimating the parameters of a sinusoid embedded in noise. The use of Kalman filters in tracking time-varying parameters in general and the phase and/or the frequency of a sinusoidal signal in particular, has been well investigated by many researchers and is extensively documented in the literature [2]-[4]. In this particular application, the parameter of primary interest is the frequency of a pure sine wave which should be estimated in the presence of high dynamics at a low carrier-to-noise ratio (CNR). Typically, it is applied in a region where the Phase Locked Loop (PLL) loses lock frequently due to cycle slipping [5] and hence is unreliable.

The scheme sought should be easy to implement and possess the same order of complexity as the PLL. The goal here is to be able to track the frequency at a low CNR with a simple scheme that trades complexity for root-mean-squared (rms) frequency error performance. The automatic frequency control (AFC) algorithms discussed in the literature are typically more complex than the PLL [6], [7] except for the cross-product AFC loop, which performs poorly in terms of lowest operating CNR. In a typical operating environment, if the PLL loses lock due to a severe maneuver by the transmitter, the proposed simple AFC algorithm could then be used to track the doppler until the dynamics are well within the tracking capability of the PLL, which can then track again with the frequency initialization provided by the AFC algorithm.

II. Description and Analysis of Algorithm

One approach to tracking the frequency of a sinusoidal signal at a lower CNR than the PLL lies in modeling the dynamic process in differential form and hence tracking only the differential dynamics. Thus in this model, the phase process can not be tracked directly and is really compromised in order to further lower the threshold at which the frequency process can be estimated. However, given its initial value, the phase can later be derived from the frequency by integration, assuming the latter is estimated with a suitably small rms error.

There are numerous ways to arrive at a differential signal model to track the frequency; for example, each of the in-phase and quadrature samples at a specific time can be expanded using a Taylor series as a linear combination of the two previous samples [8] and that, after some further processing, can lead to a signal model in the desired form. The disadvantage of this scheme is that it tends to be quite computationally demanding because it considers higher-order approximations to the measurement function. In this article, the phase information is removed by manipulating the samples in a more direct manner, i.e., a simple cross-product between the in-phase and quadrature samples is performed in a manner similar to the cross-product AFC loop. The modified samples are then fed into an Extended Kalman Filter (EKF) that tracks the differential phase change from which the doppler can be deduced.

In the derivation to follow, the vector notation which is standard in the theory of Kalman filters is adopted. The received signal is observed in the presence of additive narrowband white gaussian noise with one-sided power spectral density N_0 (watts/Hz). The carrier is first removed by mixing the observed waveform with a fixed reference, after which the in-phase and quadrature signals are sampled at a fixed rate $f_s = T_s^{-1}$ where T_s is the sampling time. Hence the samples that need to be processed can be expressed in vector form as

$$\underline{r}(k) \triangleq \begin{bmatrix} r_I(k) \\ r_Q(k) \end{bmatrix} = \begin{bmatrix} A \sin\theta(k) \\ A \cos\theta(k) \end{bmatrix} + \underline{n}(k) \qquad (1)$$

where k is the discrete time, $\theta(k)$ the phase of the received signal, A its amplitude and $\underline{n}^T(k) = [n_I(k) \;\; n_Q(k)]$ is a zero mean gaussian vector, the subscripts I and Q denoting the in-phase and quadrature components respectively. Hence,

$$E[\underline{n}(k)] = \underline{0}$$

$$E[\underline{n}(k)\underline{n}^T(k)] = \left(\frac{N_0}{2T_s}\right)\underline{\underline{I}} \qquad (2)$$

where $\underline{\underline{I}}$ denotes the 2×2 identity matrix. Typically, the received samples $r_I(k)$, $r_Q(k)$ form the input to an EKF with a state vector consisting of the phase $\theta(k)$ and its various derivatives since the latter is modeled as a polynomial in time of sufficient degree to account for the dynamics encountered. The advantage of this approach is that it tracks the phase and frequency with small rms error when operating above threshold, defined here as the CNR at which frequency loss-of-lock occurs with a probability of 0.1. Because frequency is the primary parameter in this application, we expect to further lower the operating threshold by compromising the phase estimates and the rms error performance. The block diagram of the new scheme, referred to as Frequency EKF (FEKF), is shown in Fig. 1. The cross-product is performed in order to remove the phase from the samples, as follows

$$\begin{aligned} z_I(k) &= r_I(k)\,r_Q(k-1) - r_Q(k)\,r_I(k-1) \\ z_Q(k) &= r_I(k)\,r_I(k-1) + r_Q(k)\,r_Q(k-1) \end{aligned} \qquad (3)$$

In doing so, the effective noise has been increased and this will of course result in a larger frequency error. The modified samples $z_I(k)$ and $z_Q(k)$ are then used in the EKF which tracks only the differential phase, not the pseudo-phase. This results in a reduction in the order of the EKF and hence in complexity. Defining CNR as the carrier power to the one-sided power spectral density level of the noise, we have

$$CNR = \frac{A^2}{N_0} \qquad (4)$$

From Eq. (2), N_0 is equal to $2T_s\sigma^2$ where σ^2 is the variance of the noise samples $n_I(k)$, $n_Q(k)$. Hence, CNR can be expressed in terms of σ^2 as

$$CNR = \frac{A^2}{2T_s\sigma^2} \qquad (5)$$

Without loss in generality, we can now set A to unity, as the CNR and the sampled noise variance are inversely related by Eq. (5). Plugging Eq. (1) into Eq. (3) and expanding, we easily obtain

$$\begin{aligned} z_I(k) &= \sin\Delta\theta(k) + n_I'(k) \\ z_Q(k) &= \cos\Delta\theta(k) + n_Q'(k) \end{aligned} \qquad (6)$$

where the effective noises $n_I'(k), n_Q'(k)$ are given by

$$\begin{aligned} n_I'(k) &= n_I(k-1)\sin\theta(k) + n_Q(k)\cos\theta(k-1) \\ &\quad + n_I(k-1)n_Q(k) - n_Q(k-1)\cos\theta(k) \\ &\quad - n_I(k)\sin\theta(k-1) - n_I(k)n_Q(k-1) \\ n_Q'(k) &= n_Q(k-1)\sin\theta(k) + n_Q(k)\sin\theta(k-1) \\ &\quad + n_Q(k)n_Q(k-1) + n_Q(k-1)\cos\theta(k) \\ &\quad + n_I(k)\cos\theta(k-1) + n_I(k)n_I(k-1) \end{aligned} \tag{7}$$

and the differential phase at time k is defined by

$$\Delta\theta(k) = \theta(k) - \theta(k-1) \tag{8}$$

The differential phase itself can be modeled as an nth order polynomial whose derivatives constitute the components of the state vector $\underline{x}(k)$ of the FEKF as follows

$$\begin{aligned} \Delta\theta(k) &= \underline{\ell}^T \underline{x}(k) \quad ; \quad \underline{\ell}^T = [1, 0, \ldots, 0] \\ \underline{x}(k+1) &= \underline{\underline{\Phi}}\,\underline{x}(k) + \underline{v}(k) \end{aligned} \tag{9}$$

where $\underline{v}(k)$ is a disturbance term that models the random changes in parameters due to dynamics, and $\underline{\underline{\Phi}}$ denotes the state transition matrix. In the remainder of the paper, we will concentrate on the second-order FEKF, but the analysis can be generalized to any order. For a second-order FEKF, the state $\underline{x}^T(k)$ becomes

$$\underline{x}^T(k) = \left[\Delta\theta(k) \qquad \Delta\omega(k)\right] \tag{10}$$

where $\Delta\omega(k)$ is the derivative of $\Delta\theta(k)$. In that case, Eq. (9) yields

$$\begin{aligned} \Delta\theta(k+1) &= \Delta\theta(k) + T_s\Delta\theta(k) + v_1(k) \\ \Delta\omega(k+1) &= \Delta\omega(k) + v_2(k) \end{aligned} \tag{11}$$

where

$$v_i(k) = \int_{(k-1)T_s}^{kT_s} \frac{\tau^{2-i}}{(2-i)!} J(\tau)\, d\tau \quad ; \quad i = 1, 2 \tag{12}$$

In the above, $J(t)$ stands for jerk and denotes the second derivative of the differential phase or the third derivative of the pseudo-phase. Assuming that the jerk is a zero-mean white process with one-sided spectral level N_J, we obtain

$$E[v_2^2(k)] = \frac{N_J}{2} T_s = \sigma_J^2 T_s^2 \tag{13}$$

where σ_J^2 denotes the variance of the sampled version of $J(t)$. Denoting by $\underline{\underline{Q}}$ the covariance matrix of $\underline{v}(k)$, it is easily shown [3] that

$$\underline{\underline{Q}} = \sigma_J^2 T_s^2 \begin{bmatrix} \frac{T_s^2}{3} & \frac{T_s}{2} \\ \frac{T_s}{2} & 1 \end{bmatrix} \qquad \underline{\underline{\Phi}} = \begin{bmatrix} 1 & T_s \\ 0 & 1 \end{bmatrix} \tag{14}$$

The EKF equations are derived in [1] and are repeated below for convenience in recursive form.

$$\hat{\underline{x}}(k+1/k) = \underline{\underline{\Phi}}\;\hat{\underline{x}}(k/k-1) + \underline{\underline{K}}(k)\left[\underline{z}(k) - \underline{h}(\hat{\underline{x}}(k/k-1))\right] \tag{15a}$$

$$\underline{\underline{K}}(k) = \underline{\underline{\Phi}} \sum\!\!\!\!\sum (k/k-1)\underline{\underline{H}}(k)\left(\underline{\underline{H}}^T(k)\sum\!\!\!\!\sum (k/k-1)\underline{\underline{H}}(k) + \underline{\underline{R}}\right)^{-1} \tag{15b}$$

$$\begin{aligned} \sum\!\!\!\!\sum (k+1/k) &= \alpha^2 \underline{\underline{\Phi}}\left[\sum\!\!\!\!\sum (k/k-1) - \sum\!\!\!\!\sum (k/k-1)\underline{\underline{H}}(k)\right. \\ &\quad \cdot \left(\underline{\underline{H}}^T(k)\sum\!\!\!\!\sum (k/k-1)\underline{\underline{H}}(k) + \underline{\underline{R}}\right)^{-1} \\ &\quad \left. \cdot \underline{\underline{H}}^T(k)\sum\!\!\!\!\sum (k/k-1)\right]\underline{\underline{\Phi}}^T + \underline{\underline{Q}} \end{aligned} \tag{15c}$$

where

$$\underline{h}(\underline{x}(k)) = \begin{bmatrix} \sin(\underline{\ell}^T \underline{x}(k)) \\ \cos(\underline{\ell}^T \underline{x}(k)) \end{bmatrix}$$

$$\underline{\underline{H}}^T(k) = \frac{\partial}{\partial \underline{x}} \underline{h}(\underline{x})\Big|_{\underline{x}=\hat{\underline{x}}(k/k-1)} = \begin{bmatrix} \cos\Delta\hat{\theta}(k/k-1) & 0 \\ -\sin\Delta\hat{\theta}(k/k-1) & 0 \end{bmatrix} \tag{16}$$

and $\underline{\underline{R}}$ is the covariance matrix of the noise vector $\underline{n}'(k)$ given by

$$\underline{\underline{R}} = \sigma_n^2 \underline{\underline{I}} \quad ; \quad \sigma_n^2 = 2(\sigma^2 + \sigma^4) \tag{17}$$

The weighting coefficient α is typically used to adjust the classical trade-off between adaptation time in transient situations and steady-state error. The noise sequences $n_I'(k)$, $n_Q'(k)$ are colored but Eq. (17) reflects the statistics only at a specific time. No attempt is taken to whiten the sequences in order to abide with the original goal of a simple scheme. The weighting coefficient α can be used in addition to σ_J^2 to control the effective bandwidth of the EKF, as will be shown later in the linear analysis.

The performance of the FEKF operating in the presence of noise can only be derived in the steady-state, in which case the matrix $\underline{\underline{\Sigma}}(k+1/k)$ is independent of k. From Eq. (15c), it is not clear that the FEKF will reach steady state because $\underline{\underline{\Sigma}}(k+1/k)$ depends on the matrix $\underline{\underline{H}}(k)$ which in turn depends on the predicted value $\Delta\hat{\theta}(k/k-1)$. However, it is shown in the Appendix that the right-hand side of Eq. (15c) is in fact independent of $\Delta\hat{\theta}(k/k-1)$, and that both the linear and nonlinear filters reach the identical steady state. Furthermore, an equivalent steady-state nonlinear model is derived in the Appendix which is used to compute the error variance in white noise. Defining the steady-state matrix $\overline{\Sigma}$ to be

$$\overline{\Sigma} = \lim_{k\to\infty} \underline{\underline{\Sigma}}(k+1/k) \triangleq \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix} \tag{18}$$

the equivalent linear loop model is shown in Fig. 2 where $n_{eq}(k)$ denotes the equivalent noise given by

$$n_{eq}(k) = n_I'(k)\cos\Delta\hat{\theta}(k) - n_Q'(k)\sin\Delta\hat{\theta}(k) \tag{19a}$$

with power spectrum (assuming $\phi(k) = 0$)

$$S_{n_{eq}}(z) \triangleq -\sigma^2 z + 2(\sigma^2 + \sigma^4) - \sigma^2 z^{-1} \tag{19b}$$

Letting $\phi(k) \triangleq \Delta\theta(k) - \Delta\hat{\theta}(k)$ denote the differential error phase at time k, it is shown in the Appendix that the error variance is given by

$$\sigma_\phi^2 = T_s \int_{-\frac{1}{2T_s}}^{\frac{1}{2T_s}} \left| H\left(e^{js\pi fT_s}\right)\right|^2 S_{n_{eq}}\left(e^{j2\pi fT_s}\right) df \tag{20}$$

where $H(z)$ denotes the closed loop transfer function, related to the loop filter $F(z)$ via

$$H(z) \triangleq \frac{F(z)}{1+F(z)} = \frac{z(\sigma_1^2 + T_s\rho) - \sigma_1^2}{z^2(\sigma_1^2 + \sigma_n^2) + z(T_s\rho - \sigma_1^2 - 2\sigma_n^2) + \sigma_n^2} \tag{21}$$

The closed loop transfer function is depicted in Fig. 3 for a fixed α (1.005), N_J (200), and T_s(2 msec). The corresponding loop bandwidth B_L defined in the Appendix is shown in Fig. 4 versus the ratio σ_J^2/σ_n^2. When the measurement noise is dominant (i.e., $\sigma_J^2/\sigma_n^2 \ll 1$), the bandwidth is independent of CNR and is controlled by α. On the other hand, when the loop is dominated by the process noise (i.e., $\sigma_J^2/\sigma_n^2 \gg 1$), the bandwidth is a function of both CNR and α. The performance in the absence of dynamics is shown in Fig. 5 as a function of CNR when T_s = 2 msec. The theory and simulation are in agreement as long as the loop signal-to-noise ratio (SNR) is "high." For low loop SNR, the loop is nonlinear and the performance is degraded. Note from Fig. 6 that the noise power spectrum is not always white in the loop bandwidth, hence Eq. (20) can not be simplified any further in general.

III. Performance in a Dynamic Environment

The performance of the FEKF in a dynamic environment can only be assessed through simulations. In order to compare the performance with other schemes, the FEKF was tested in the presence of the identical dynamics described in [9], which exhibit two 100 g/sec jerks lasting for 0.5 sec each. For a fixed sampling time of 2 msec, the best performance in terms of lowest achievable CNR threshold is obtained when α is 1.005 and N_J is equal to 300. This corresponds to 22.5 dB-Hz (Fig. 7) with an rms frequency error 41.2 Hz (Fig. 8). The loop bandwidth at threshold is about 7.1 Hz with a 35.2 Hz steady-state error due to the jerk. The contribution of the noise in the linear model is about 27.3 Hz. Note that for fixed α and N_J, the loop bandwidth is a function of CNR as mentioned earlier.

Compared with other frequency tracking schemes using the same trajectory, the threshold of the FEKF is 3.5 dB lower than the PLL, 2.2 dB better than the CPAFC loop, 1.5 dB better than a fourth-order EKF tracking phase, and finally 0.5 dB more efficient than the approximate Maximum Likelihood (ML) scheme described in [9]. However in terms of rms frequency error, the FEKF is worse than all the above loops except for the CPAFC loop which exhibits an inferior performance. In terms of complexity, the FEKF requires the same number of computations per update as the PLL when implemented in the steady state.

IV. Conclusion

A new AFC loop was introduced and analyzed. The heart of the loop involves an EKF which operates on a modified set of data in order to track the frequency. The scheme can also be viewed as a modification of the well known cross-product AFC loop.

A detailed analysis of the second-order loop was presented and verified via simulations. The parameters of the FEKF were related to tracking parameters such as loop bandwidth and frequency jitter due to noise. The steady-state error due to jerk was also assessed.

The algorithm is best suited for scenarios in which frequency error is of secondary value and lowest operating CNR threshold is of primary concern. This technique is easily implemented and requires a minimum amount of computations per update. Moreover, it is highly suited for the ARX as it requires minimum software changes.

References

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, New Jersey: Prentice Hall, 1979.

[2] D. R. Polk and S. C. Gupta, "Quasi-Optimum Digital Phase Locked Loops," *IEEE Trans. Comm.*, pp. 75–82, January 1973.

[3] F. R. Castella, "An Adaptive Two-Dimensional Kalman Tracking Filter," *IEEE Trans. Aero. Elec. Sys.*, vol. AES-16, no. 6, pp. 822–829, November 1973.

[4] B. Friedland, "Optimum Steady-State Positions and Velocity Estimation Using Noisy Sampled Position Data," *IEEE Trans. Aero. Elec. Sys.*, vol. 9, pp. 906–911, November 1973.

[5] W. C. Lindsey, *Synchronization Systems in Communication and Control*, New Jersey: Prentice Hall, 1972.

[6] F. D. Natali, "AFC Tracking Algorithms," *IEEE Trans. Comm.*, vol. COM-32, no. 8, pp. 935–947, August 1984.

[7] W. J. Hurd, J. I. Statman, and V. A. Vilnrotter, "High Dynamic GPS Receiver Using Maximum Likelihood Estimation and Frequency Tracking," *IEEE Trans. Aero. Elec. Sys.*, vol. AES-23, no. 4, pp. 925–937, July 1987.

[8] R. Kumar, "Differential Sampling for Fast Frequency Acquisition Via Adaptive Extended Least Squares Algorithm," *Proceedings of the International Telemetering Conference*, San Diego, California, pp. 191–201, October 1987.

[9] V. A. Vilnrotter, S. Hinedi, R. Kumar, *A Comparison of Frequency Estimation Techniques for High Dynamic Trajectories*, JPL Publication 88-21, Jet Propulsion Laboratory, Pasadena, California, September 15, 1988.

[10] B. Ekstrand, "Analytical Steady State Solution for a Kalman Tracking Filter," *IEEE Trans. Aero. Elec. Sys.*, vol. AES-19, no. 6, pp. 815–819, November 1983.

[11] E. I. Jury, *Theory and Application of the Z-Transform Method*, New York: Wiley, 1964.

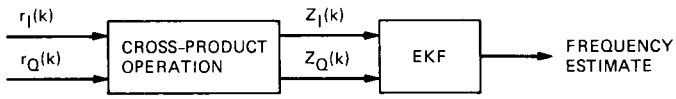


Fig. 1. General block diagram of the FEKF.

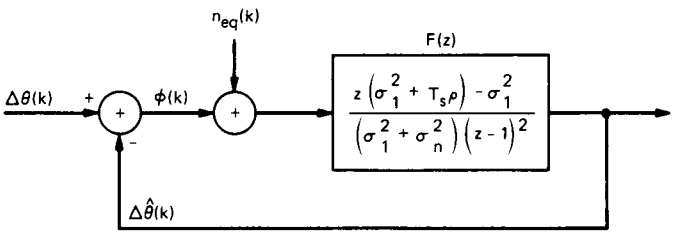


Fig. 2. Equivalent linear model of the second-order FEKF.

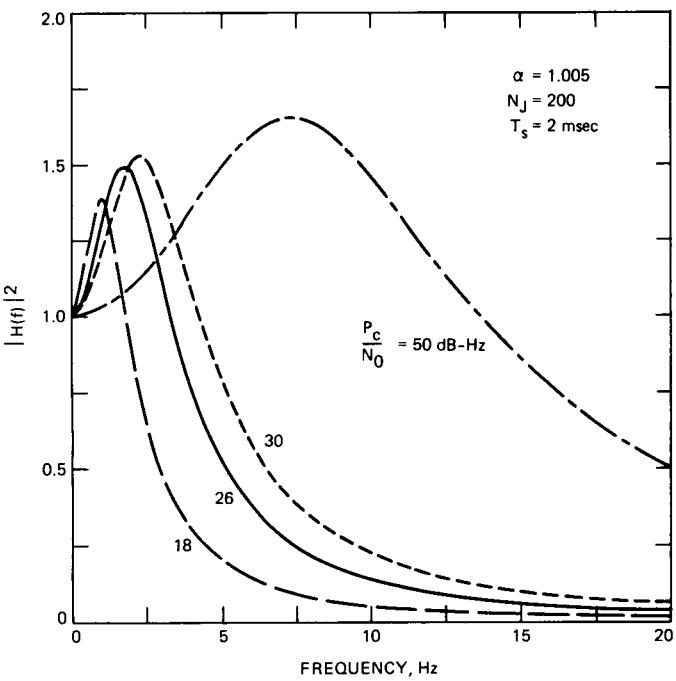


Fig. 3. Closed loop transfer function.

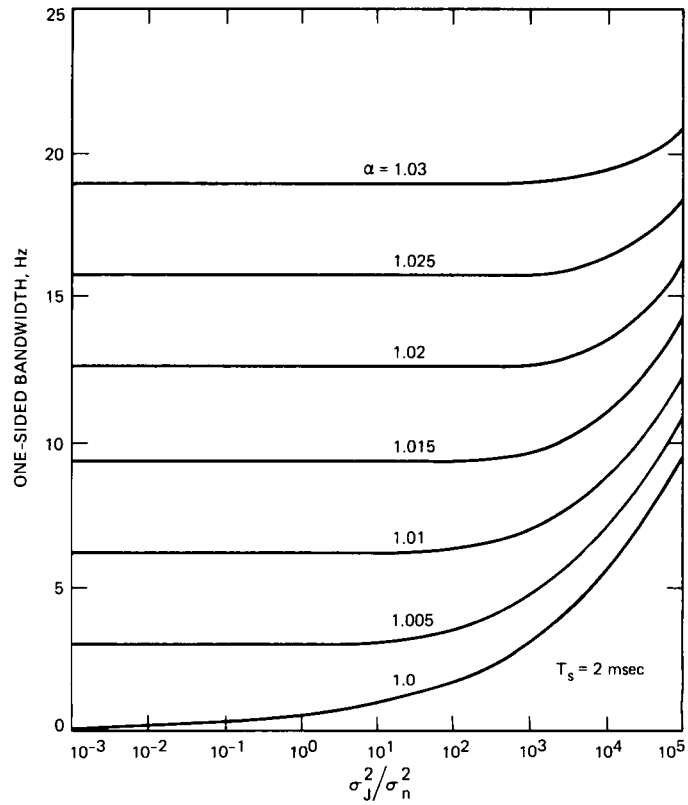


Fig. 4. Loop bandwidth.

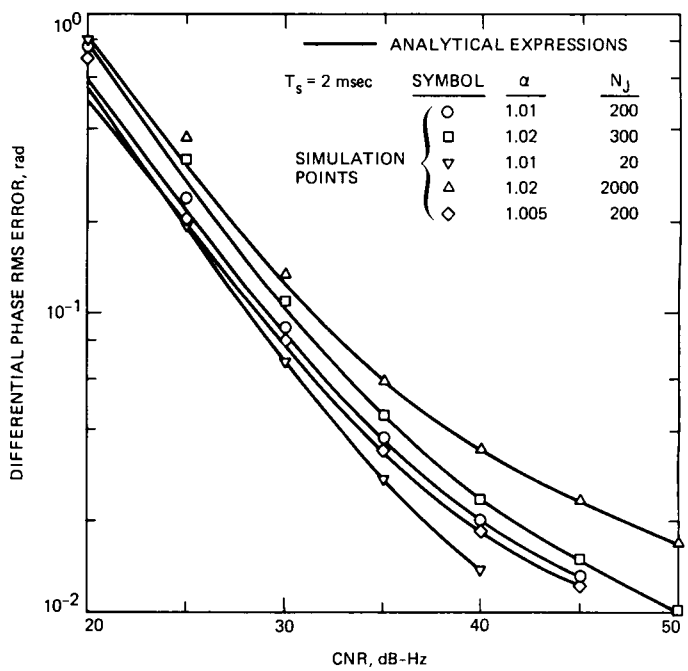


Fig. 5. Differential phase error versus CNR.

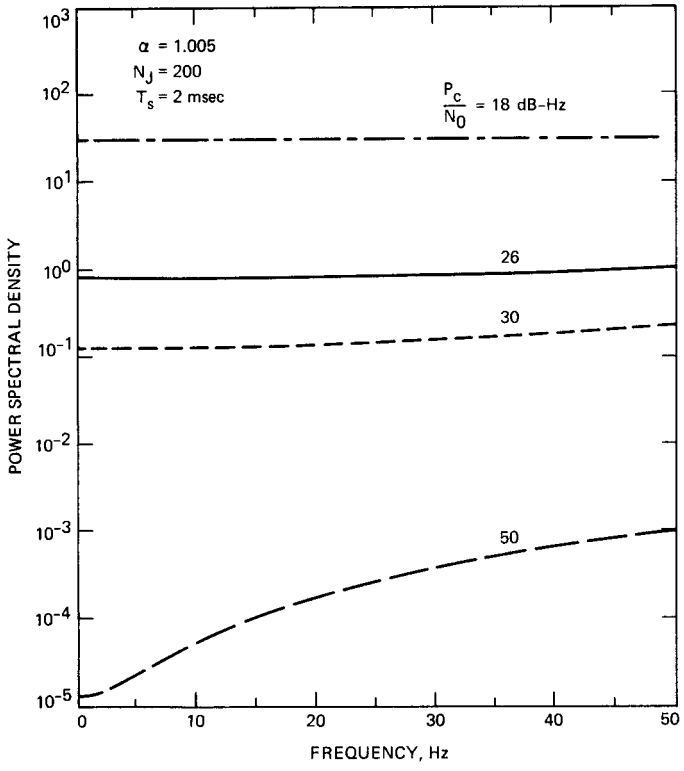


Fig. 6. Equivalent noise power spectral density.

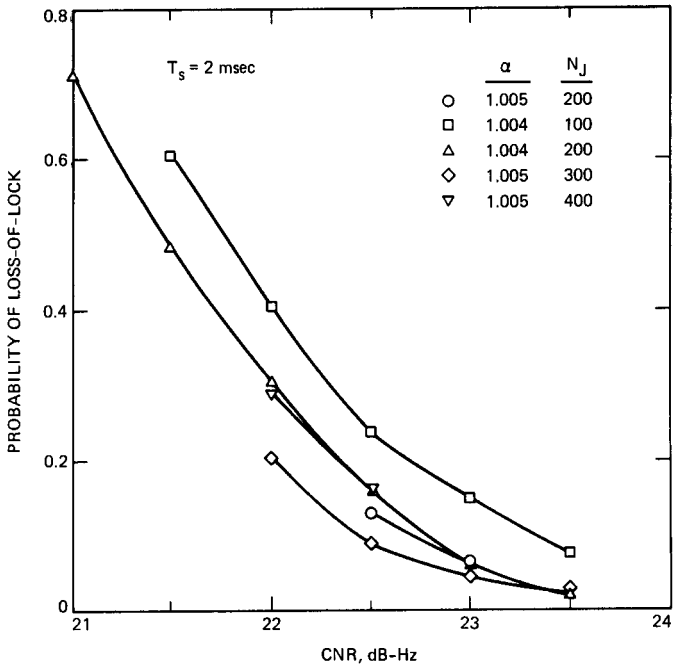


Fig. 7. Probability of loss-of-lock versus CNR.

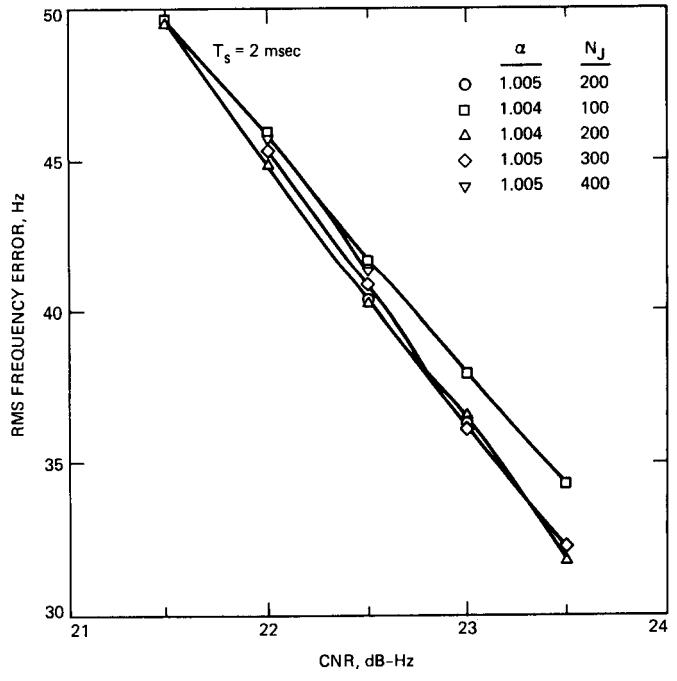


Fig. 8. RMS frequency error versus CNR.

Appendix
Performance of the FEKF in the Steady State

The error covariance matrix $\underline{\underline{\Sigma}}(k+1/k)$ satisfies the recursive Eq. (15c) which seems to depend on $\Delta\hat{\theta}(k/k-1)$. Defining the matrix

$$\underline{\underline{\Omega}}(k) = \underline{\underline{H}}^T(k)\,\underline{\underline{\Sigma}}(k/k-1)\,\underline{\underline{H}}(k) + \underline{\underline{R}} \tag{A-1}$$

and replacing $\underline{\underline{\Sigma}}(k/k-1)$ by its steady-state value given by Eq. (18), we have (letting $\underline{x}$ denote $\Delta\hat{\theta}(k/k-1)$)

$$\underline{\underline{\Omega}} = \begin{bmatrix} \cos x & 0 \\ -\sin x & 0 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \cos x & -\sin x \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \sigma_n^2 & 0 \\ 0 & \sigma_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 \cos^2 x + \sigma_n^2 & -\sigma_1^2 \cos x \sin x \\ -\sigma_1^2 \cos x \sin x & \sigma_1^2 \sin^2 x + \sigma_n^2 \end{bmatrix} \tag{A-2}$$

Inverting $\underline{\underline{\Omega}}(k)$, we obtain

$$\underline{\underline{\Omega}}^{-1}(k) = \frac{1}{\sigma_n^2\,\sigma_1^2 + \sigma_n^4} \begin{bmatrix} \sigma_1^2 \sin^2 x + \sigma_n^2 & \sigma_1^2 \sin x \cos x \\ \sigma_1^2 \cos x \sin x & \sigma_1^2 \cos^2 x + \sigma_n^2 \end{bmatrix} \tag{A-3}$$

which when pre- and post-multiplied by $\underline{\underline{H}}(k)$ gives

$$\underline{\underline{H}}(k)\,\underline{\underline{\Omega}}^{-1}(k)\,\underline{\underline{H}}^T(k) = \begin{bmatrix} \cos x & -\sin x \\ 0 & 0 \end{bmatrix} \cdot \underline{\underline{\Omega}}^{-1}(k) \begin{bmatrix} \cos x & 0 \\ -\sin x & 0 \end{bmatrix}$$

$$= \frac{1}{\sigma_1^2 + \sigma_n^2} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{A-4}$$

which is independent of $\Delta\hat{\theta}(k/k-1)$. The steady-state matrix $\overline{\Sigma}$ can be computed in closed form only when α is equal to one [10], otherwise the solution can only be obtained numerically. One approach is to run the FEKF until it reaches steady state and the resulting matrix is a solution of Eq. (15c) because of the properties of Kalman filters. Note also that the linear measurement (i.e., $z(k) = \Delta\theta(k) + n'(k)$) with $\underline{H}^T = [1 \;\; 0]$ would have resulted in the same matrix as in Eq. (A-4) and hence in the identical matrix $\overline{\Sigma}$.

In the steady state, the measurement vector $z(k)$ is subtracted from the prediction $\underline{h}(\cdot)$ before being multiplied by the gain matrix $\underline{\underline{K}}(k)$ as in Eq. (15a). Defining $\underline{\underline{L}}(k)$ to be

$$\underline{\underline{L}}(k) = \sum\!\!\!\!\!\!\sum (k/k-1)\,\underline{\underline{H}}(k)\,\underline{\underline{\Omega}}^{-1}(k) \tag{A-5}$$

we obtain in the steady state ($x \triangleq \Delta\hat{\theta}(k/k-1)$)

$$\underline{\underline{L}}(k) = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \cos x & -\sin x \\ 0 & 0 \end{bmatrix} \underline{\underline{\Omega}}^{-1}(k)$$

$$= \frac{1}{\sigma_n^2 + \sigma_n^2} \begin{bmatrix} \sigma_1^2 \cos x & -\sigma_1^2 \sin x \\ \rho \cos x & -\rho \sin x \end{bmatrix} \tag{A-6}$$

which when multiplied by $\underline{z}(k) - \underline{h}(\cdot)$ yields

$$\underline{\underline{L}}(k)\,[\underline{z}(k) - h(\hat{\underline{x}}(k/k-1))] = \underline{\underline{L}}(k) \begin{bmatrix} z_I(k) - \sin\Delta\hat{\theta}(k/k-1) \\ z_Q(k) - \cos\Delta\hat{\theta}(k/k-1) \end{bmatrix}$$

$$= [z_I(k)\cos\Delta\hat{\theta}(k/k-1) - z_Q(k)\sin\Delta\hat{\theta}(k/k-1)] \cdot \frac{1}{\sigma_1^2 + \sigma_n^2} \begin{bmatrix} \sigma_1^2 \\ \rho \end{bmatrix} \tag{A-7}$$

Equation (A-7) combined with Eq. (15a) suggests the structure depicted in Fig. A-1. The output of the frequency disciminator $d(k)$ is given by

$$d(k) = z_I(k)\sin\Delta\hat{\theta}(k/k-1) - z_q(k)\cos\Delta\hat{\theta}(k/k-1) \tag{A-8}$$

Using Eq. (6), $d(k)$ simplifies to

$$d(k) = \sin\phi(k) + n_{eq}(k) \tag{A-9}$$

where $q(k)$ denotes the differential phase error and

$$n_{eq}(k) = n_I'(k)\cos\Delta\hat{\theta}(k/k-1) - n_q'(k)\sin\Delta\hat{\theta}(k/k-1) \tag{A-10}$$

It is straightforward to show that

$$\begin{aligned} E[n_{eq}^2(k)] &= 2(\sigma^2+\sigma^4) \\ E[n_{eq}(k)\,n_{eq}(k\pm1)] &= -\sigma^2\cos[\phi(k)-\phi(k-1)] \end{aligned} \tag{A-11}$$

which, assuming zero error, results in the following power spectrum

$$S_{n_{eq}}(z) = -\sigma^2 z + 2(\sigma^2+\sigma^4) - \sigma^2 z^{-1} \tag{A-12}$$

The resulting nonlinear model is shown in Fig. A-2. Using z-transforms, it is straightforward to show that

$$\underline{\hat{X}}(z) = \frac{1}{(z-1)^2}\begin{bmatrix} z-1 & T_s \\ 0 & z-1 \end{bmatrix}\underline{W}(z) \tag{A-13}$$

where $\underline{\hat{X}}(z)$ and $\underline{W}(z)$ are the $\underline{z}$-transforms of $\underline{\hat{x}}(k/k-1)$ and $\underline{\omega}(k)$, respectively. The loop filter $F(z)$ is then easily derived and is equal to

$$\begin{aligned} F(z) &= [1\ \ 0]\frac{1}{(z-1)^2}\begin{bmatrix} z-1 & T_s \\ 0 & z-1 \end{bmatrix}\begin{bmatrix} 1 & T_s \\ 0 & 1 \end{bmatrix} \\ &\quad\cdot\frac{1}{(\sigma_1^2+\sigma_n^2)}\begin{bmatrix}\sigma_1^2 \\ \rho\end{bmatrix} \\ &= \frac{z(\sigma_1^2+T_s)-\sigma_1^2}{(\sigma_1^2+\sigma_n^2)(z-1)^2} \end{aligned} \tag{A-14}$$

Figure A-3 depicts the simplified nonlinear model in terms of the loop filter $F(z)$. Approximating $\sin(\phi(k))$ by $\phi(k)$ for small error, we have, using operator notation

$$\phi(z) = [1-H(z)]\,\Delta\theta(z) - H(z)\{n_{eq}(k)\} \tag{A-15}$$

where $H(z)$ is the closed loop transfer function given by

$$H(z) = \frac{F(z)}{1+F(z)} = \frac{z(\sigma_1^2+T_s\rho)-\sigma_1^2}{z^2(\sigma_1^2+\sigma_n^2)+z(T_s\rho-\sigma_1^2-2\sigma_n^2)+\sigma_n^2} \tag{A-16}$$

From classical digital phase locked loop analysis, it is straightforward to show that the error in the absence of dynamics is given by

$$\sigma_\phi^2 = T_s\int_{\frac{-1}{2T_s}}^{\frac{1}{2T_s}}\left|H\left(e^{j2\pi fT_s}\right)\right|^2 S_{n_{eq}}\left(e^{j2\pi fT_s}\right)df \tag{A-17}$$

while the steady-state error due to jerk (in units of m/sec^3) is

$$\phi_{ss}(rad) = \left(\frac{\omega_i}{c}\right)\frac{J_0T_s^2}{\rho}(\sigma_1^2+\sigma_n^2) \tag{A-18}$$

where ω_i is the radian frequency of the incoming signal and c the velocity of light (in m/sec). The one-sided closed loop bandwidth $B_L(Hz)$ defined by

$$B_L(Hz) = \frac{1}{2T_s}\frac{1}{H^2(1)}\frac{1}{2\pi i}\oint H(z)\,H(z^{-1})\frac{dz}{z} \tag{A-19}$$

can be computed in closed form using the results found in [11] to give

$$B_L(Hz) = \frac{1}{2T_s}\frac{B_0e_1 - B_1a_1}{(a_0^2-a_2^2)(a_0+a_2)-(a_0-a_2)a_1^2}$$

where

$$\left.\begin{aligned} b_2 &= -\sigma_1^2;\ b_1 = \sigma_1^2+T_s\rho \\ a_2 &= \sigma_n^2;\ a_1 = T_s\rho-\sigma_1^2-2\sigma_n^2 \\ a_0 &= \sigma_1^2+\sigma_n^2 \end{aligned}\right\} \tag{A-20}$$

and

$$\left.\begin{aligned} B_0 &= b_1^2+b_2^2;\ B_1 = 2b_1b_2 \\ e_1 &= a_0+a_2 \end{aligned}\right\} \tag{A-21}$$

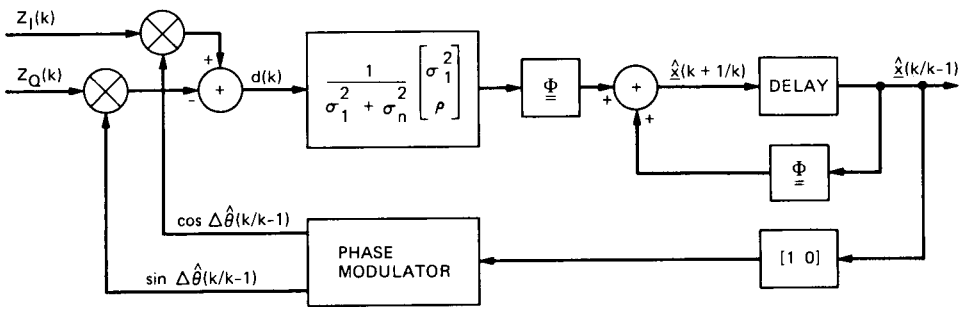


Fig. A-1. Equivalent steady-state model.

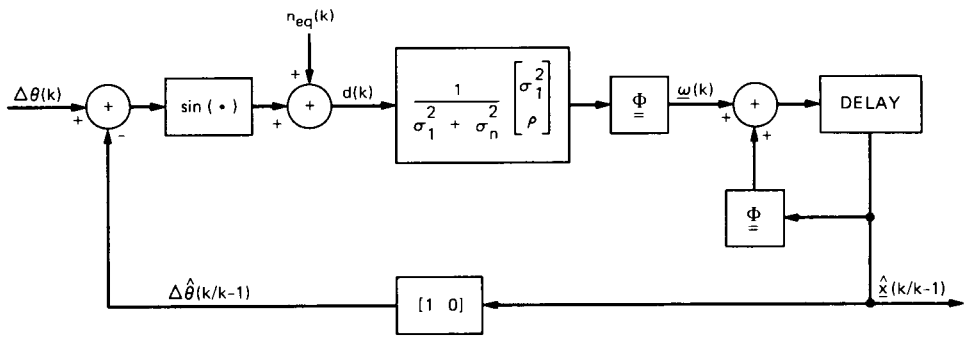


Fig. A-2. Equivalent nonlinear model.

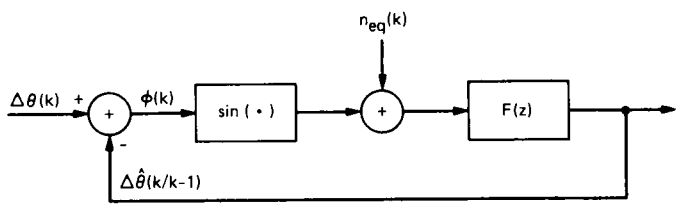


Fig. A-3. Simplified nonlinear model.

Transmitter Data Collection Using Ada

B. L. Conroy
Radio Frequency and Microwave Subsystems Section

This article describes a data collection system installed on the 400-kilowatt X-band transmitter of the Goldstone Solar System Radar. The data collection system is built around off-the-shelf IEEE 488 instrumentation, linked with fiber optics, controlled by an inexpensive computer, and uses software written in the Ada language. The speed and accuracy of the system is discussed, along with programming techniques used for both data collection and reduction.

I. Introduction

The system described in this article is the result of two separate goals. The first goal was to instrument the Goldstone Solar System Radar (GSSR) transmitter to start building a base of operating data that could be used for statistical analysis. The second goal was to demonstrate the feasibility of a data collection system that used only readily available, easily calibrated instruments. This article is concerned with the extent to which this second goal has been realized.

The Ada language was chosen for a number of features: (1) multitasking is provided within the language, (2) it embodies a number of software engineering disciplines such as modularity, strong typing, and levels of abstraction, which will result in robust and more maintainable programs, and (3) the federal government maintains a standard for the language, and enforces conformity to the standard, which will contribute to maintainability of Ada programs.

II. Hardware Description

Figure 1 is a block diagram of the entire system, as installed at DSS 14. The control computer and one data collection unit are located in Room 105. This data collection unit has been wired into the Local Control Console (LCC) to provide access to 26 analog values, 53 interlocks, 40 indicators, and 9 warnings. The second data collection unit is located in Module II, just below the tri-cone area, and connected to Room 105 by a fiber-optic link. This unit monitors 18 resistive temperature detectors (RTDs), 4 turbine flow meters, and 8 paddle-wheel flow meters, all of which were added in the radar feedcone during the 70-meter upgrade period.

A. Control Computer

An IBM industrial AT computer was chosen as the controller. This computer is similar to the standard AT, but is rack-mounted and provided with an air filter and vibration damping mounts for expansion cards. It has an IEEE 488 interface card, which allows a wide range of instruments to be monitored.

B. Data Collection Units

The HP 3852S Data Collection Unit is a rack-mounted card case, with local programming capability and a wide range of plug-in cards. The cards used in this system include a 5½-digit voltmeter (option 44701A), a 5-channel counter (option

44715A), and several multiplexer cards. Reference 1 contains a summary of cards available, and their capabilities.

C. Fiber Optic

The data collection unit in the cone area is linked to the pedestal by a pair of HP 37204A fiber-optic bus extenders. These units are completely transparent to the control computer, and use internal protocols to ensure error-free data transmission. The maximum data transmission rate of 60,000 bytes per second is adequate for this application.

III. Software Description

Figure 2 is a "Booch" diagram of the overall structure of the controlling software at a high level of abstraction. This style of diagram is introduced in [2], which is an excellent introduction to Ada. The central unit is the HARDWARE_DEFINITIONS package, which contains most of the information specific to the transmitter. This package provides a list of all the parameters available (an ENUMERATION type), functions that return the current value and status of each parameter, and an internal task that periodically reads the data collection units. Figure 3 is the declaration of the type PARAMETER, with explanatory comments. The names chosen for the parameters are abbreviations descriptive of the function. For example, Filament Time Delay on klystron 1 becomes FIL_TD_1 (an interlock) and Alidade Heat Exchanger On becomes ALI_HE_ON (a warning because the main heat exchanger should be adequate). Some other abbreviations used are UC for Under Current, CB for Crow Bar, IGN for Ignatron, IL for Interlock, RPA for Reflected Power Amplifier, COLL for Collector, V for Voltage, I for Current, and UA for Microamps.

The GSSR program contains the operator interface. It contains three display tasks, each of which can be directed to display a different view of the data, and a keyboard watching task that switches the views based on operator inputs. The DATA_LOG package monitors the measured data and writes it to disk.

A. Utility Packages

The SCREEN_IO package provides facilities for multiple screen windows, each with its own attributes. These windows are used by the display tasks to present independent views of the measured data. It also provides functions to monitor the keyboard. Although they are not used in this application, it provides control over the screen modes and graphics capability for both color graphics and enhanced graphics adapters.

The IEEE_488 package provides facilities for sending and receiving ASCII or binary data, sending bus commands, serial or parallel polling, and detecting service requests. It contains an internal task for resource control. This is needed in a multitasking environment to ensure that different tasks cannot send conflicting messages.

B. Update

The UPDATE package encapsulates all the information about the data collection unit configurations. Several techniques are used to minimize the number of transactions on the IEEE 488 bus. First is the use of downloaded subroutines in the scanners. On startup, this package loads each scanner with a program that will take all the required measurements. This allows a set of measurements to be taken with the command "CALL DOMEAS" rather than a repetition of what measurements are to be taken. The second technique is the use of block transfers in binary rather than individual data values in ASCII. This reduces the number of bytes sent by about a factor of 3. A third technique is preprocessing the interlocks and indicators (which are either 0 or 28 volts) and sending only the numbers of the channels that exceed a threshold (2 bytes) rather than the actual values (8 bytes). Fourth is overlapping the measurement time of one scanner with the data transfer time of the other scanner.

C. Calculated Parameters

In addition to the parameters measured directly, there are a number of calculated ones. Thermal techniques as described in [3] are used to calculate the power in the water loads on each of the two klystrons and the waster load on the four-port power combiner. In addition, a 4-foot section of the waveguide between the power combiner and the feedhorn has been calibrated and instrumented for a thermal determination of the total power being delivered to the antenna. The time remaining in the present cycle (transmit or receive) is calculated from the round-trip light time (entered by the operator) and the time of the last change in the beam status. This module also determines the correct scale factor for the vacuum-ion-pump current on each klystron from the range indicators.

D. Data Log

The DATA_LOG package contains a set of limits on each parameter, and a task that records all analog parameters whenever one or more of the limits is exceeded.

When the program is started, the operator is prompted for the name of the target, the round trip light time (used to calculate the TIME_LEFT parameter) and a file name for the data. The file name is passed to the LOG task, which creates the file and writes a header consisting of the data, the target, an optional smoothing factor, and the names of the parameters that will be recorded. The LOG task waits until all parameters have been measured, then writes their initial values to the data

file. One of the parameters recorded is the measurement time, which time tags the data record. The format used for recording is Comma Separated Values (CSV) that is, the ASCII representation of the values, with commas separating them.

The decision on when to record data is based on a "record on change" algorithm. The DATA_LOG package contains a limit on the absolute value of change for each measured parameter. Every time the data is updated, the LOG task compares the change in each parameter from the last value recorded. If the change on any parameter exceeds the limit for that parameter, it sets a flag. If no parameter has changed more than its limit, the data is kept in temporary storage. When a change does occur, two sets of data are recorded: the last measurement that did not show a change and the measurement that did show a change. This technique simplifies the plotting of the data.

IV. Data Reduction

Supercalc 4 is used for data reduction. One of its options is reading CSV data files, and it allows keyboard macros that can import data, set the scales and labels, and plot the data with a few keystrokes. Figure 4 is a graph of the transmitter output power during a recent experiment. This graph plots the total output power (P_TOT), along with output powers of each klystron (P_OUT_1 and P_OUT_2). Since all operating parameters are recorded, additional graphs, such as beam voltage and drive power, can be produced to determine the source of the variations in the output power. The output power can also be supplied to the scientists for calculation of target albedo or cross section.

V. Results and Discussion

A. Speed and Accuracy

The data collection system reads four turbine flow meters with 1-hertz accuracy. The actual frequencies range from about 900 hertz to 1200 hertz, so this is about 0.1 percent of the actual reading. There are 18 RTD temperature sensors, which are read to about 0.1°C accuracy at a rate of 25 readings per second. The system measures 32 dc voltages with 5½-digit precision (100 readings per second) and 120 dc voltages with 3½-digit precision (160 readings per second). Using the overlapping techniques described above, it makes a measurement of every parameter every 3.5 seconds.

The major limit on the time for a measurement cycle is the integration time of the precision voltmeter. Because one scanner is transmitting its data while the other is making measurements, an increase in the data transfer rate would not decrease the cycle time. The counter card needs 1 second of integration time to get 1-hertz accuracy, but it reads all channels simultaneously, and is independent of the voltmeter, so it is not limiting the overall speed.

If necessary, some speed improvement is possible in the future by using a higher-speed voltmeter (option 44702A) for some of the less critical measurements. This unit has only 12 bits of resolution (plus sign) and a maximum input voltage of 10 volts, but can read up to 100,000 channels per second.

B. Problems

During the initial phases of the development, efforts were made to allow the system to function with bad or missing sensors by marking individual data items as "UNKNOWN," but measuring and recording the rest of the data. This effort was only partially successful. At present, it can detect and allow for problems in the RTD temperature sensors and failures of an entire data collection unit. A problem at the card level within a data collection unit prevents reading of other cards, and a failure of the fiber-optic link prevents reading of the local data collection unit.

C. Open Items

Not yet implemented in the system are: (1) better techniques of detecting and flagging hardware failures, (2) more precise limits on parameters used in the "record on change" algorithm, (3) other tools for data reduction, and (4) real-time graphical representation of certain parameters.

References

[1] Hewlett Packard, HP 3852S, *Data Acquisition and Control System Data Book*, 1986.

[2] G. Booch, *Software Engineering with Ada*, Second Edition, Menlo Park, California: Benjamin/Cummings, 1986.

[3] B. Conroy, H. Schleier, and T. Tesarek, "Thermal Evaluation Method for Klystron RF Power," *TDA Progress Report 42-88*, October–December 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 91–95, February 15, 1987.

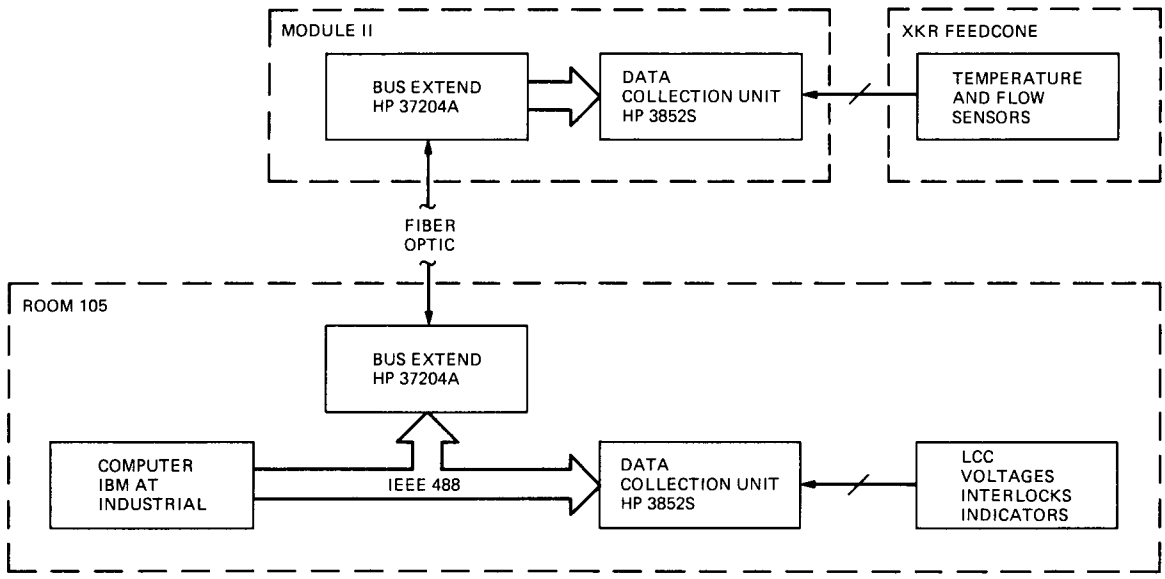


Fig. 1. Block diagram of data collection system.

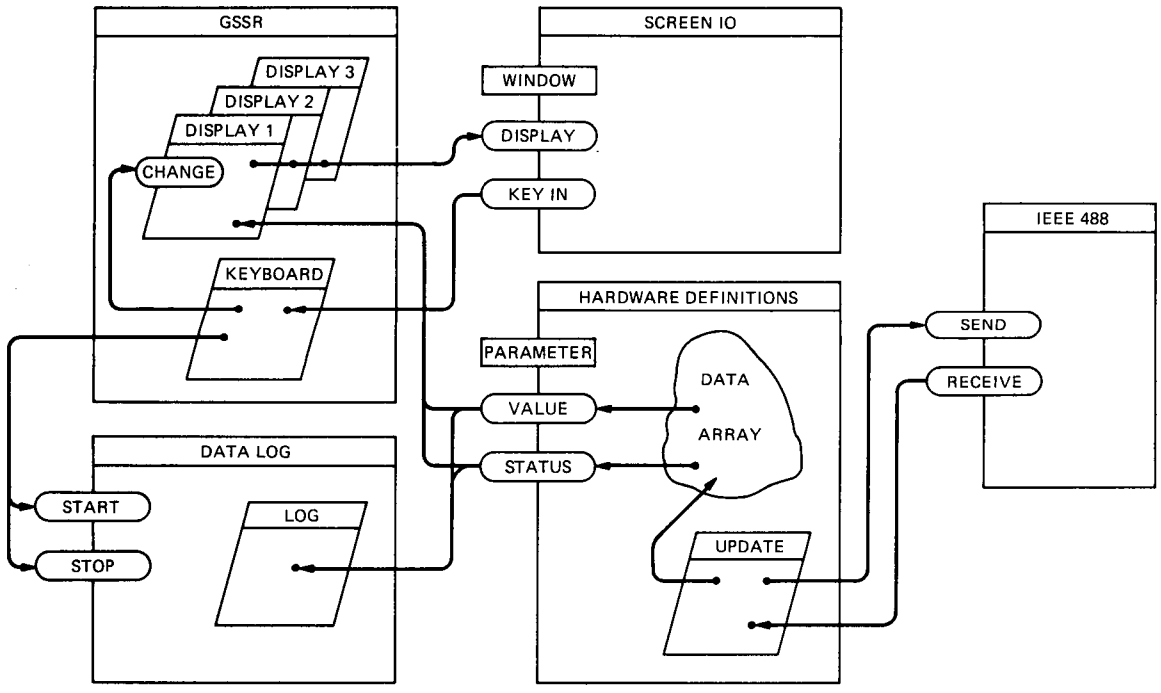


Fig. 2. "Booch" diagram of data collection system software.

```
type PARAMETER is (

-- analog parameters
   -- Klystron 1, upstairs system
   LOAD_FLOW_1, LOAD_TIN_1, LOAD_TOUT_1, LOAD_TURB_1,
   COLL_FLOW_1, COLL_TIN_1, COLL_TOUT_1,
   BODY_FLOW_1, BODY_TIN_1, BODY_TOUT_1,
   MAG_FLOW_1, WASTE_TOUT_1,
   -- Downstairs system
   COLL_I_1, FIL_I_1, FIL_V_1, MAG_I_1,
   P_OUT_1, P_DRIV_1, P_REFL_1, VAC_I_1,
   RPA_ONE_1, RPA_TWO_1,
   -- calculated parameters
   P_LOAD_1, P_WASTE_1,

   -- Klystron 2, upstairs system
   LOAD_FLOW_2, LOAD_TIN_2, LOAD_TOUT_2, LOAD_TURB_2,
   COLL_FLOW_2, COLL_TIN_2, COLL_TOUT_2,
   BODY_FLOW_2, BODY_TIN_2, BODY_TOUT_2,
   MAG_FLOW_2, WASTE_TOUT_2,
   -- Downstairs system
   COLL_I_2, FIL_I_2, FIL_V_2, MAG_I_2,
   P_OUT_2, P_DRIV_2, P_REFL_2, VAC_I_2,
   RPA_ONE_2, RPA_TWO_2,
   -- calculated parameters
   P_LOAD_2, P_WASTE_2,

   -- Common Parameters, Upstairs
   WG_TURB, WG_TIN, WG_TOUT,
   P_TOT_TURB, P_TOT_TIN, P_TOT_TOUT,
   -- Downstairs
   BODY_I, BEAM_V, BEAM_I, CB_TIME,
   VAC_V, PHASE, VDC_28, P_TOT,
   -- Misc
   P_WASTE, -- sum of two waster loads
   TIME_LEFT, -- derived from RTLT
   MEASUREMENT_TIME,
   -- calculated
   P_TOT_CALC,

-- Indicators
   -- Klystron 1
   UA_5_1, UA_50_1, UA_500_1, UA_5000_1, UA_50000_1, -- Vac Ion Scale
   FIL_RAISE_1, MAG_RAISE_1, DRIVE_RAISE_1,
   FIL_LOWER_1, MAG_LOWER_1, DRIVE_LOWER_1,
   -- Klystron 2
   UA_5_2, UA_50_2, UA_500_2, UA_5000_2, UA_50000_2, -- Vac Ion Scale
   FIL_RAISE_2, MAG_RAISE_2, DRIVE_RAISE_2,
   FIL_LOWER_2, MAG_LOWER_2, DRIVE_LOWER_2,
   -- common
   BEAM_READY, BEAM_ON, BEAM_RAISE, BEAM_LOWER,
   S_BAND_DSN, S_BAND_RADAR, X_BAND_RADAR,
   ANT_POS, LOAD_POS,
   HE_ON, MAIN_HE_ON, MG_ON, DRIVE_ON,
   PHASE_0, PHASE_180, IL_OPEN,
   PGM_MODE, COMP_IF_ON,
```

Fig. 3. Ada declaration of type PARAMETER.

```
-- Warnings
   ALI_HE_ON, AUX_HE_ON, RESIST_IN, RESIST_OUT,
   HE_TANK_LOW, HE_TANK_PRESS, HE_FANS,
   WG_PRESS, TR_FLOW,

-- Interlocks
   -- Klystron 1
   FIL_TD_1, FIL_UC_1, COLL_OC_1, FOCUS_UC_1,
   REF_PWR_ONE_1, REF_PWR_TWO_1, REF_METER_1,
   ARC_DET_ONE_1, ARC_DET_TWO_1,
   COLL_FLOW_LO_1, BODY_FLOW_LO_1,
   LOAD_FLOW_LO_1, DRIFT_FLOW_LO_1, FIL_FLOW_LO_1,
   VAC_PWR_1, CB_MAG_1,
   -- Klystron 2
   FIL_TD_2, FIL_UC_2, COLL_OC_2, FOCUS_UC_2,
   REF_PWR_ONE_2, REF_PWR_TWO_2, REF_METER_2,
   ARC_DET_ONE_2, ARC_DET_TWO_2,
   COLL_FLOW_LO_2, BODY_FLOW_LO_2,
   LOAD_FLOW_LO_2, DRIFT_FLOW_LO_2, FIL_FLOW_LO_2,
   VAC_PWR_2, CB_MAG_2,
   -- common
   ELEVATION, PS_DOOR, PA_DOOR, CB_DOOR,
   TXR_CONFIG, MICROWAVE, CB_TEST, CB_FIRED,
   FAST_BODY, SLOW_BODY, IGN_PWR, BODY_OC,
   DC_OV, DC_OC, PHASE_FAIL, TR_OIL_LOW,
   HV_ZERO, MOTOR_ST, MOTOR_LO, GEN_LO, ALI_HE_OT,

-- other parameters
   OUTPUT_TO, CONFIG,
   TARGET, TIME, DATE, UNUSED, NONE);
```

Fig. 3 (contd).

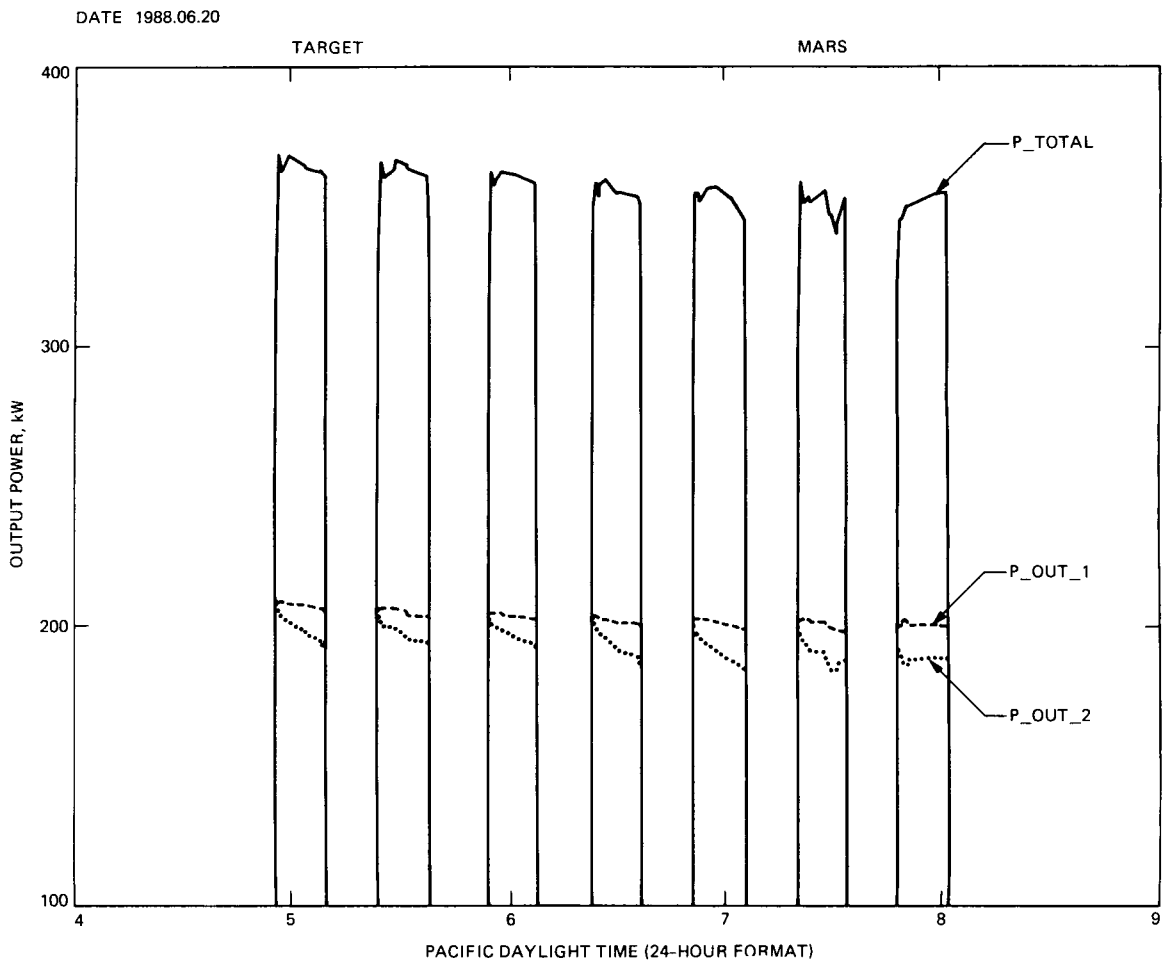


Fig. 4. Typical data plot.

DSN 70-Meter Antenna X-Band Gain, Phase, and Pointing Performance, With Particular Application for Voyager 2 Neptune Encounter

S. D. Slobin and D. A. Bathker
Radio Frequency and Antenna Microwave Subsystems Section

The gain, phase, and pointing performance of the DSN 70-m antennas are investigated using theoretical antenna analysis computer programs that consider the gravity-induced deformation of the antenna surface and quadripod structure. The microwave effects are calculated for normal subreflector focusing motion and for special fixed-subreflector conditions that may be used during the Voyager 2 Neptune encounter. The frequency stability effects of stepwise lateral and axial subreflector motions are also described. Comparisons with recently measured antenna efficiency and subreflector motion tests are presented. A modification to the existing 70-m antenna pointing squint correction constant is proposed.

I. Introduction

With the approaching Voyager 2 Neptune encounter, scheduled for August 25, 1989, it is important to characterize the operation of the DSN 70-m antenna network in its newly upgraded configuration. It was predicted that the upgrade would improve antenna gain by 55 percent (1.9 dB) or more, relative to the gain in its 64-m configuration. This has been accomplished and verified for the three 70-m antennas by efficiency measurements using radio sources as reference standards. The gain/noise temperature ratio has been increased even more due to a decrease of system noise temperature at all elevation angles. A more complete description of these results will be reported in future TDA Progress Reports.

Normally, antenna subreflector (SR) focusing motion is made to optimize the gain at all elevation angles. This subreflector motion is necessary because the antenna changes shape and the quadripod moves (a total of three inches laterally and one-half inch axially) as a function of elevation angle. Because the best-fit focus of the deformed main reflector is generally not coincident with the subreflector focus (after quadripod movement), subreflector movement commanded by the subreflector controller (SRC) is necessary. Although the antenna gain is optimized by this method, pointing is not preserved. A "squint correction" to the predicted elevation angle for SR lateral motion along the antenna Y-axis (vertical for AZ-EL antennas pointing at the horizon) is made by offsetting the predicted elevation angle as a function of subreflector position only (not elevation angle). Somewhat more complete descriptions of these antenna and subreflector motions for the 64-m antennas are given in [1] and [2]. The study in [1] grew out of a problem that occurred during the Voyager 2 Uranus encounter wherein a mispositioned subreflector degraded the

gain and pointing capabilities of the DSS-43 (Australia) 64-m antenna. A recent study [3] compares 70-m antenna performance predicted by both GTD (Geometrical Theory of Diffraction) calculations and traditional ray tracing methods.

For the purposes of frequency stability of the DSS-43 70-m antenna during the Voyager 2 Neptune closest approach (approximately 0700 to 0915 GMT, Earth-received time, over an elevation angle range of approximately 45 to 70 degrees), two modifications of subreflector positioning are being considered: (1) fix the subreflector in its axial (Z) motion at a position corresponding to an elevation angle of approximately 60 degrees, or (2) fix the subreflector in both its axial (Z) and vertical (Y) motions at a position corresponding to an elevation angle of 60 degrees. The subreflector X-axis position (horizontal) does not change, as there is no net gravity component in that direction, and hence no net change in the loading as a function of elevation angle. A previous study concerning the phase and frequency stability of Cassegrainian antennas (e.g., DSS-43 in its 64-m configuration and the present DSS-42 HA-DEC antenna) is presented in [4]. Typical maximum subreflector axial (Z) rates of motion are 0.050 in./sec, resulting in an RF path length change of 0.090 in./sec. (In the study reported in [4] it was found that the ratio of change in RF path length to change in subreflector axial position was 1.8. (An earlier study [5] determined this factor to be 1.76.) For X-band frequency (8420 MHz), this results in a frequency signature (for this example) of about 0.064 Hz. This low frequency signature confuses radio science data analysis, hence the proposal to fix subreflector motion. The purpose of the study presented here is to determine the gain and phase effects of this nonstandard antenna configuration. It should be noted that the DSN 70-m antennas have somewhat different subreflector velocities than those used for the 64-m antennas. The maximum z-rate is not typically used. Actual operational methods will be discussed in Section III.

Figure 1 shows the elevation angles of the three 70-m antennas on the day of Voyager 2 Neptune encounter. The proposed tracking plan at DSS-43 entails normal tracking using conscan and normal subreflector Y-Z focusing until an elevation angle of about 40-43 degrees is reached. The subreflector would then be moved abruptly in Z, fixed in its 60-degree position, and conscan would be turned off. (At the time of the writing of this report, the Voyager 2 radio science plan is to fix the subreflector motion in Z-axis only.) The spacecraft would be tracked in the fixed-subreflector mode up to about 70-degree elevation (covering the period between the two 71K-Ring Earth Occultation events). At an elevation of about 70 degrees, the subreflector would be returned to automatic mode and conscan turned on for the remainder of the pass except for a short interval during Triton occultation.

The structural model of the DSN 70-m antenna surface was developed by JPL's Ground Antenna and Facilities Engineering Section (R. Levy and M. S. Katow, private communication). For reference, the model is designated JRMFL-FX-03J/70M-MAR88, and at this time (August 1988) is the most recent model being used to describe antenna structural deformation as a function of gravity loading. The subreflector controller (SRC) model (P. Lipsius, private communication) used in the calculations here is given by equations describing subreflector motions along the Y (lateral) and Z (axial) directions:

$$Y = -0.0369\,A - 6.4516\,B + 0.0257$$

$$Z = -1.7270\,A - 0.0732\,B - 0.0982$$

where

$A = \sin(45^\circ) - \sin(\mathrm{EL})$

$B = \cos(45^\circ) - \cos(\mathrm{EL})$

EL = elevation angle

These equations are programmed in PROMs in the SRC. Additional SR position offsets are computed and applied in field operation by numerous measurements made to maximize efficiency at a particular elevation angle (D. Girdner and W. Wood, Goldstone Deep Space Communications Complex, private communication). Typically, both best antenna shape and maximum efficiency are obtained at 45-degree elevation, and bias terms (different from SR position offsets) arise from uncontrollable nonzero values occurring in installation of position readouts, etc. All 70-m antennas have identical PROMs in the SRC containing identical constant terms, however operator-applied inputs from the Local Monitor and Control (LMC) console could vary from antenna to antenna as a result of differing local conditions. Errors in panel setting, later structural modification, or special cases may result in changes from the 45-degree optimum position. For the calculations performed here, it is assumed that perfect shape and maximum gain are obtained at 45-degree elevation. For these GTD calculations only, the constants in the above equations describing subreflector position were made equal to zero.

It should be noted that in the GTD computer programs currently used to calculate dual-shaped reflector performance, the description of subreflector and feed relative positions requires special attention. It is important to note that the net subreflector movement in the main reflector coordinate system is a combination of both the quadripod movement and the subreflector movement relative to the quadripod. As the feedhorn is referenced to the subreflector coordinate system, yet remains

fixed (to a good approximation) in the main reflector coordinate system, its position relative to the "moving" subreflector coordinate system must also be calculated.

II. Computed Gain Effects

Figure 2 shows the GTD-computed gains for the three subreflector modes of operation: (1) operating normally in automatic mode, (2) Z fixed in its 60-degree elevation position, and (3) Y and Z fixed in their 60-degree elevation positions. The computed gains include only the losses due to aperture illumination, feed and subreflector spillover, phase error, and cross-polarization. Effects that are not included are waveguide and dichroic plate loss, quadripod blockage, VSWR, surface roughness, and several other items. The predicted theoretical operational antenna gain is about 1 dB lower than the values shown in Fig. 2. Table 1 shows all the components going into the design expectation of operational gain (at the peak gain elevation angle of 45 degrees). Note that the design expectation of 74.39 dBi (71.94-percent efficiency) is 0.96 dB below the GTD-calculated value of 75.35 dBi (89.74-percent efficiency). The actual measured antenna efficiencies for the 70-m network are about 68 percent (peak, without atmosphere, at a 45-degree elevation angle), corresponding to an antenna gain of 74.15 dBi. The quarter-dB difference between design expectation and actual performance will be investigated to test the validity of the values presented in Table 1. For this study the absolute values of gain and efficiency are not critical to the phase-effect problem. It is the difference between the normal automatic mode of SR operation and the two fixed-SR modes that gives the effect import for radio science data.

It is seen in Fig. 2 that the Z-fixed SR condition shows less than 0.1-dB loss of gain over the elevation range of 45 to 70 degrees. The actual Z-mispositioning of the subreflector remains small over that elevation angle range, compared to the mispositioning resulting from fixing Y also. For example, at 45-degree elevation, the subreflector is mispositioned 0.259 inches in Z and 1.330 inches in Y, relative to their positions at a 60-degree elevation angle. From previous studies,[1] the loss arising from the Z-mispositioning is predicted to be

$$\Delta G = 2.18(0.259)^2 \text{ dB}$$

$$= 0.146 \text{ dB}$$

This loss is somewhat larger than the 0.1 dB shown in Fig. 2.

For the case of Y and Z fixed, there also exists a mispositioning of the subreflector in Y (at elevation angles differing from 60 degrees), and this mispositioning results in about 1.5 beamwidths of scan at 45-degree elevation. Pointing effects will be discussed later. From the previously quoted study by R. Levy, the gain loss at 45 degrees associated with the 1.330-in. mispositioning of the subreflector in Y is given by

$$\Delta G = 0.28(1.330)^2 \text{ dB}$$

$$= 0.495 \text{ dB}$$

This is almost exactly the difference between the Z-fixed and Y/Z-fixed curves in Fig. 2. It appears then that the R. Levy model may overestimate the effect of the Z-mispositioning of the subreflector. A further comparison of the R. Levy model and the GTD-calculated gain losses should be made.

For preservation of gain, it is seen that fixing the subreflector Z-movement at its 60-degree position has minimal effect over the 45- to 70-degree elevation Neptune encounter period.

III. Computed Phase Effects

Figure 3 shows the far-field phase effect associated with normal SR movement. There is a total of more than 650 degrees of phase change over the elevation range of 5 to 90 degrees. This phase change is due almost entirely to the Z-movement of the subreflector.

A model for the particular curve shown in Fig. 3 is

$$\text{phase} = 814 - 718 \sin(\text{EL})$$

where EL = elevation angle, degrees.

Note that the constant term in the above equation is arbitrary, and could be set to zero. What is important is the phase variation with elevation angle.

The net Z subreflector movement over this range is 1.048 in. (1.504 in. commanded by the SRC and -0.456 in. from quadripod movement) relative to the main reflector vertex, and includes both the effect of the SR movement and the deformation of the quadripod structure. Tests carried out at Goldstone[2] and subsequent GTD calculations (described later in this report) indicate that the effect of Z subreflector

[1] R. Levy, "Gain Losses for Non-Optimal Antenna Subreflector Offsets," JPL IOM 3325-88-009 (internal document), February 5, 1988.

[2] R. Riggs, "Preliminary Report of the Results of the Radio Science Tests Conducted at DSS-14," JPL IOM RLR-88-10 (internal document), May 10, 1988.

movement for a 70-m antenna (K-band feed in the XKR cone) is to change the path length by about 1.76 times the amount of movement. For Y subreflector movement using the same feedhorn, the path length factor was determined to be 0.093. A 1.048-in. net subreflector Z-movement thus gives 1.844 inches of path length change, or about 474 degrees of phase at 8420 MHz. The additional phase change is due to the deformation of the main reflector itself. The frequency signature of this phase change if the SR movement were absolutely smooth and continuous is probably of little consequence for radio science investigations. For an elevation change from 45 to 50 degrees over a time interval of 0.410 hours, the phase change is 42.4 degrees, resulting in a frequency signature of about 0.08 mHz. The actual Y and Z SR movements are made in abrupt steps, resulting in a much higher frequency signature.

Figure 4 shows a comparison among the normal, fixed-Z, and fixed-Y/Z subreflector conditions. Note that it is the fixing of the SR Z-movement that predominantly gives rise to the much reduced phase change. The residual change of phase results from both the quadripod movement and main reflector deformation. These deformations work in opposite directions as far as path length changes are concerned, and hence the net phase change effect is small.

In contrast to the smoothly changing phase effects experienced when both Y and Z subreflector positions are fixed, the actual operational phase changes occur "abruptly" when the subreflector position is changed in small steps. For the 70-m antenna, the subreflector position is updated when an error of more than 0.009 in. is detected in either the Y or Z positions. The subreflector is commanded to move until the position error becomes less than 0.007 in. With overshoot due to motor and gear system inertia, the total movement is thought to be 0.005 to 0.006 in. (K. Nikbakht, private communication). The maximum sustained rate that the subreflector is capable of moving is 1 in./min (0.017 in./sec) in the Y direction, and 3.14 in./min (0.052 in./sec) in the Z direction. The subreflector is not generally operated in this mode, however, as the mechanical stresses are quite large. In actuality, gradual motor acceleration and reduced maximum speed result in the subreflector movements over the 0.005–0.006 in. range occurring in approximately 2 sec. The average rate resulting is thus about 0.003 in./sec, substantially lower than the Y and Z maximum rates. It will be shown in Section V that for the XRO cone on all 70-m antennas, the GTD-computed path length change for Y-movement is 0.0444 times the subreflector movement and for Z-movement is 1.671 times the subreflector movement. (The XRO cone is located at the upper left cone position when looking into the face of the main reflector.) Thus, the average pathlength rates for Y and Z, respectively, are about 0.00013 in./sec and 0.0050 in./sec. For a frequency of 8420 MHz, this results in Y and Z doppler frequency signatures of about 0.093 MHz and 3.6 MHz, respectively, with a probable uncertainty of at least 30 percent.

IV. Computed Pointing Effects

Figure 5 shows the actual antenna beam pointing resulting from the three different subreflector positioning schemes. In the case of the normal automatic SR movement, it is seen that the beam moves upward somewhat faster than the increasing elevation angle. Indeed, from a 45- to 80-degree elevation angle, the beam has moved up an additional 142 millidegrees as a result of the attempt to maintain optimum gain by proper positioning of the subreflector. In order to maintain proper pointing, the applied squint correction is the negative of this beam movement.

It can be seen from Fig. 5 that over the elevation range of 45 to 70 degrees, the Z-fixed subreflector condition results in negligible additional pointing error, to within the resolution of these calculations (1 millidegree). A simple analysis [6], [7] shows that even with the asymmetric off-axis tri-cone structure of the 70-m antennas, the 0.259-in. mispositioning of Z (for the Z-fixed subreflector condition) at 45-degree elevation results in a pointing error of less than 0.5 millidegrees. The reason for this is that Z-errors are just in-out position changes of the subreflector along the main reflector Z-axis. Fixing Y results in a deviation from normal pointing. The squint correction operates according to the Y-position of the subreflector relative to the quadripod structure, not from the net subreflector position, which includes quadripod movement also. It will be shown in Section V that the current (August 1988) squint correction as implemented in the Antenna Controller Subsystem (ACS) does not accurately follow the curve shown in Fig. 5. It has been necessary to employ a systematic error correction table to correct for such discrepancies. For the subreflector to be both Y- and Z-position fixed, the squint correction will be maintained constant over the 45- to 70-degree elevation range, and substantial pointing errors will result. For radio science purposes, it appears that fixing only Z introduces no effect that will seriously compromise the pointing capabilities of the antenna. The systematic pointing error correction developed with the SR in normal automatic mode should then be correct for SR Z-fixed operations.

V. Effects of Lateral and Axial Subreflector Motion at a 45-degree Elevation Angle and Modifications to Existing Squint Correction

A series of GTD calculations was made at 8420 MHz to assess the effects of lateral and axial subreflector motions on gain, phase, and pointing for both the X-band (XRO) and

K-band cones on the 70-m antennas. Looking into the antenna face with the antenna pointing at the horizon (elevation angle equal to 0 degrees), the K-band cone is at the bottom (6-o'clock, 0-degree clock angle) and the X-band cone is at the upper left (10-o'clock, 120-degree clock angle). Table 2 presents the results of these calculations for the X-band cone with the antenna positioned at a 45-degree elevation angle (and hence with a perfect surface). It is seen that for the movements shown, substantial (tenths of a dB) gain changes result. For Y-subreflector movements, small phase changes result, whereas for Z-movements, substantial phase changes occur. From these calculations, dimensionless "K-factors" can be derived linking phase change to subreflector movement (inches of path length change/inches of subreflector movement). Thus phase change (degrees) at any frequency can be determined. For the X-band (XRO) cone at a clock angle of 120 degrees, the values of K determined were:

$$K_y = +0.0444$$

$$K_z = -1.671$$

These are consistent with the propagation sign convention, $\exp(-jkr)$, in the GTD program where increasing path length results in more negative (or decreasing) phase.

For the K-band cone, where Y-movements of the subreflector are directly "toward" or "away" from the cone, a surprising result occurs. Although the subreflector movement is "lateral" in both cases, the value for the K-band cone turns out to be

$$K_y = -0.0878$$

This value is almost exactly double the value (but of opposite sign) for the X-band cone. The K-band K_y and K_z compare well with the experimentally measured values of -0.093 and -1.76, respectively.

For subreflector movements purely lateral (at 90 degrees) to the cone direction, it is found that

$$K_y = 0.00078$$

which is deemed (for the purposes of this report) to be zero. (The nonzero result is undoubtedly due to roundoff error in the calculations.)

It is postulated that the value of K (lateral) is proportional to both the distance of the feed from the main reflector axis and to the component of subreflector movement directly "toward" or "away" from the feed in question. Thus, for a feed located directly at the center of the main reflector, very small subreflector movements would result in negligible phase changes. Figure 6 shows the lateral K-factor as a function of clock angle for cones located on three subreflector focus circles of varying diameter. The diameters increase in size from R_1 to R_3 ($R_1 = 0.25R_2$, $R_2 = 0.5R_3$). The curve with maximum amplitude (R_3) is the K-factor curve for cones located along the 70-m antenna subreflector focus circle (radius equal to 108.03 cm). Phase effects determined from these curves may assist in future positioning of outrigger horns or additional cones on the existing 70-m antennas.

The existing 70-meter squint correction, as implemented in the Antenna Controller Subsystem (ACS), uses the subreflector Y-position multiplied by a constant (0.0342 degrees/inch of SR Y-axis position) to calculate the amount of correction to the elevation angle needed to maintain the beam on target during a track. The existing constant as determined in [3] was verified in this study using similar computational methods. However, this constant predicts the amount of beam movement due to subreflector motion only; it does not account for additional beam movement arising from quadripod motion and main reflector deformation as the elevation angle changes. It is found from this study that another constant more appropriately predicts beam movement, using the SR Y-position as an *indicator* of total antenna "condition," rather than as the entire cause of beam mispointing.

Figure 7 shows the actual GTD-calculated beam-peak offset as a function of normal SR Y-position. It is seen that at a 90-degree elevation angle, more than 30 millidegrees of pointing error exist. This is larger than the 70-m 3-dB beamwidth! From this curve a new squint correction constant may be found:

$$K_{squint} = 0.04145 \text{ degrees/inch}$$

where the SRC Y-position (inches) is used as the indicator of squint correction needed. This new constant could be installed in the ACS as a replacement for the existing constant. (If this is implemented, new systematic pointing-error correction tables would have to be developed.)

It should be noted that the use of this new constant during normal subreflector focusing tests will now result in mispointing of the beam, as the original squint constant (0.0342) was really the correct one to use for conditions of subreflector movement only—it was not appropriate for normal tracking operations.

A new squint correction may be generated from the current squint correction plus a correction as a function of elevation angle (Fig. 8). The new squint correction as a function of SR Y-position (Y_{sr}) and elevation angle (EL) is given by

$$\text{SQUINT}(Y_{sr}, \text{EL}) = 0.0342\,(Y_{sr}) - 0.0329 + 0.0468\cos(\text{EL})$$

If this new squint correction is implemented, the result will be a "hybrid" squint correction model using both the old constant as a function of SR position and an additional pointing correction as a function of elevation angle. The additional pointing correction will also be generated in the ACS.

VI. Conclusion

The extensive series of calculations described here indicate several methods by which the Voyager Radio Science team might avoid the confusing effects of rapidly changing phase in the August 1989 Neptune encounter data. It appears that fixing the subreflector Z-motion over the elevation range of 45 to 70 degrees adequately solves this problem without significant gain degradation. A useful result of the gain, phase, and pointing determinations is a possible modification to the existing squint correction used on the 70-m antennas. This modification substantially changes the value of the "squint constant" in the ACS and appears to more accurately model existing antenna pointing as a function of subreflector position during normal tracking operations.

Acknowledgment

The authors wish to acknowledge the assistance of R. L. Riggs of the TDA Engineering Office. His technical guidance has aided in making this study of immediate value to the Deep Space Network and the Voyager Project.

References

[1] S. D. Slobin and W. A. Imbriale, "DSS-43 Antenna Gain Analysis for Voyager Uranus Encounter: 8.45-GHz Radio Science Data Correction," *TDA Progress Report 42-90*, vol. April–June 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 127–135, August 15, 1987.

[2] C. N. Guiar and L. W. Duff, "64-M Antenna Automatic Subreflector Focusing Controller," *TDA Progress Report 42-78*, vol. April–June 1984, Jet Propulsion Laboratory, Pasadena, California, pp. 73–78, August 15, 1984.

[3] J. M. Schredder, "Seventy-Meter Antenna Performance Predictions: GTD Analysis Compared With Traditional Ray-Tracing Methods," *TDA Progress Report 42–92*, vol. October–December 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 166–174, February 15, 1988.

[4] A. G. Cha, "Phase and Frequency Stability of Cassegrainian Antennas," *Radio Science*, vol. 22, no. 1, pp. 156–166, January–February 1987.

[5] T. Y. Otoshi and W. V. T. Rusch, "Multipath Effects on the Time Delay of Microwave Cassegrain Antennas," *DSN Progress Report 42–50*, Jet Propulsion Laboratory, Pasadena, California, pp. 52–55, January–February 1979.

[6] A. M. Isber, "Obtaining Beam-Pointing Accuracy with Cassegrain Antennas," *Microwaves*, vol. 6, pp. 40–44, August 1967.

[7] Y. T. Lo, "On the Beam Deviation Factor of a Parabolic Reflector," *IRE Trans. Antennas and Propagat.*, vol. AP-8, pp. 347-349, May 1960.

Table 1. Design expectations for 70-m antenna with stovepipe feedhorn at 8420 MHz and 45-degree elevation angle

Item	Loss, dB	Net gain, dBi
100% area efficiency		75.82
GTD-included losses illumination amplitude illumination phase forward and rear spillover subreflector blockage m ≠ 1 modes cross-polarization	−0.47	75.35
Waveguide loss	−0.07	
Dichroic plate loss	−0.035	
VSWR	−0.039	
Quadripod blockage	−0.454	
Antenna surfaces (0.7-mm rms) main reflector panel mfg. main reflector panel setting subreflector surface	−0.192	
Stovepipe feed compromise	−0.11	
Imperfect focus alignment	−0.05	
Panel gaps	−0.01	74.39
	(= 71.94% efficiency)	

Table 2. Effects of lateral and axial subreflector movement for X-band cone (clock angle = 120 degrees)

Displacements		GTD-computed gain, dBi at 8420 MHz	GTD-computed phase, deg at 8420 MHz
Y, inches	Z, inches		
+1.0	0	75.07	119.6
+0.5	0	75.28	114.2
0.0	0	75.35	108.5
−0.5	0	75.28	102.7
−1.0	0	75.07	97.0
0	+0.30	75.24	−20
0	+0.15	75.35	44
0	0.00	75.35	108.5
0	−0.15	75.26	173
0	−0.30	75.06	237

Results:

K_y = +0.0444

K_z = −1.671

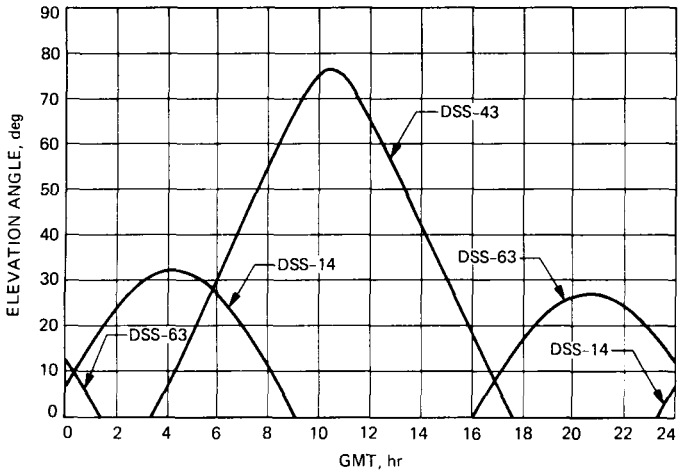


Fig. 1. Antenna elevation angles at Neptune Encounter, August 25, 1989, Earth-receive time.

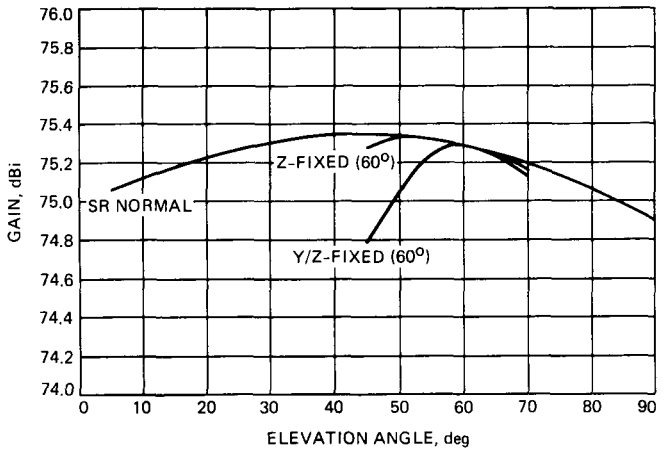


***Fig. 2. 70-m antenna GTD-computed gain versus elevation* angle, 8420 MHz, with three different subreflector configurations.**

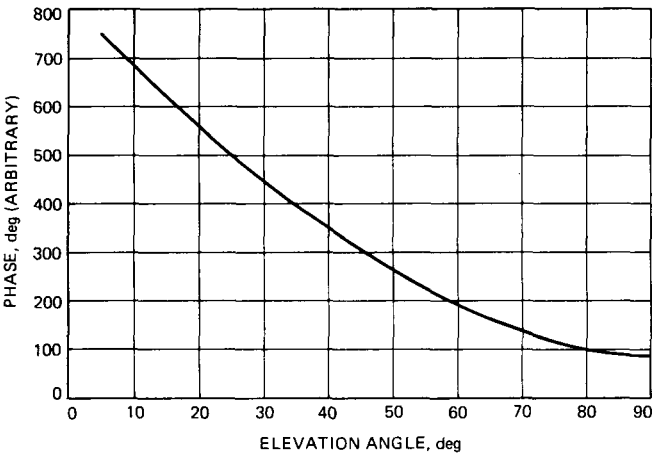


Fig. 3. 70-m antenna GTD-computed phase versus elevation angle, 8420 MHz, SR normal operation, arbitrary phase values.

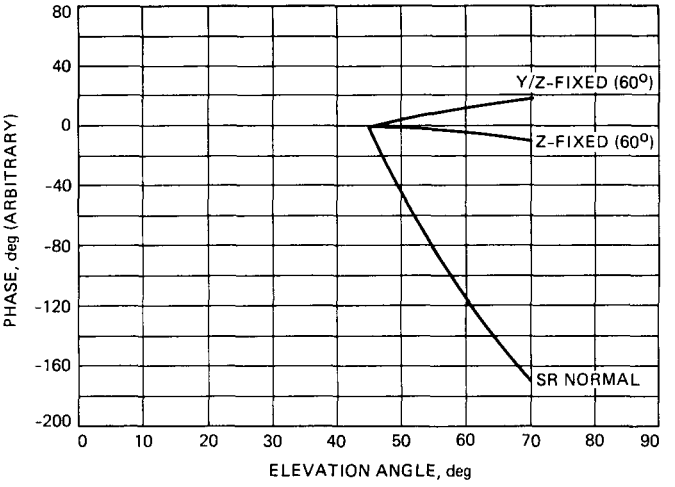


Fig. 4. 70-m antenna GTD-computed phase versus elevation angle, 8420 MHz, for three different subreflector configurations.

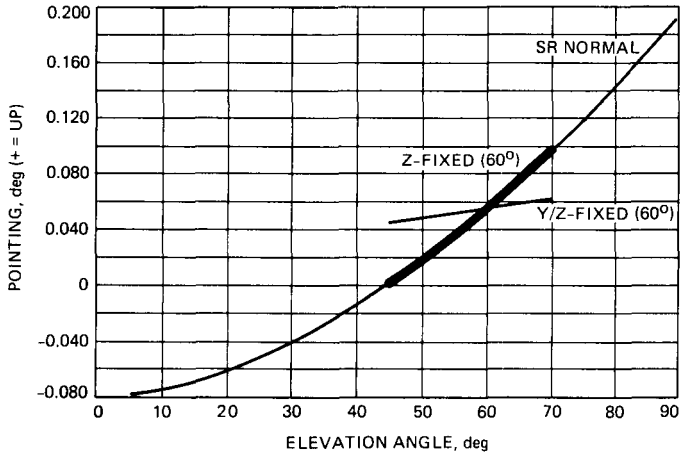


Fig. 5. 70-m antenna GTD-computed pointing versus elevation angle, 8420 MHz, for three different subreflector configurations.

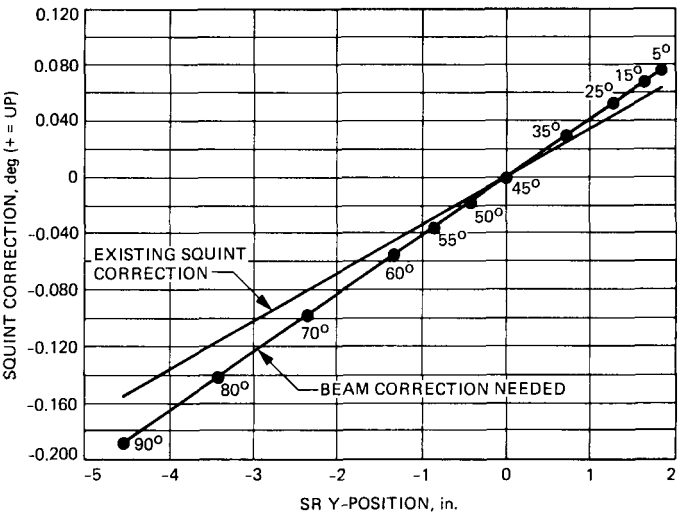


Fig. 7. 70-m antenna squint correction versus SR Y-position, existing squint correction and needed beam correction, elevation angle indicated.

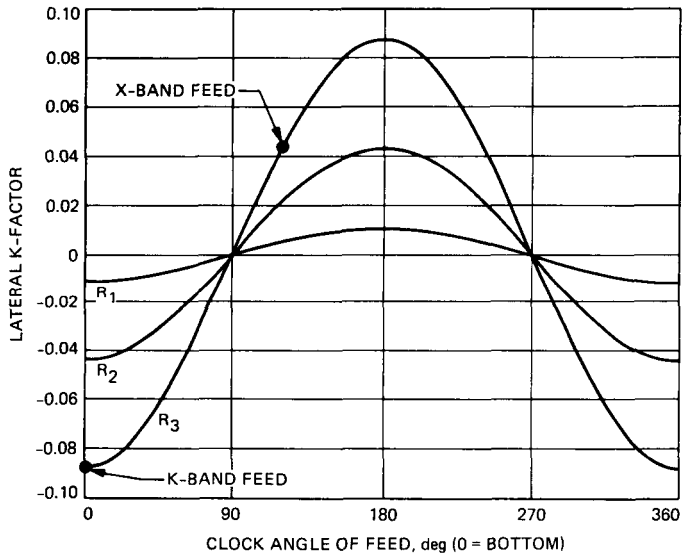


Fig. 6. Subreflector lateral movement K-factor as a function of feed clock-angle and radial distance from 70-m antenna main reflector axis.

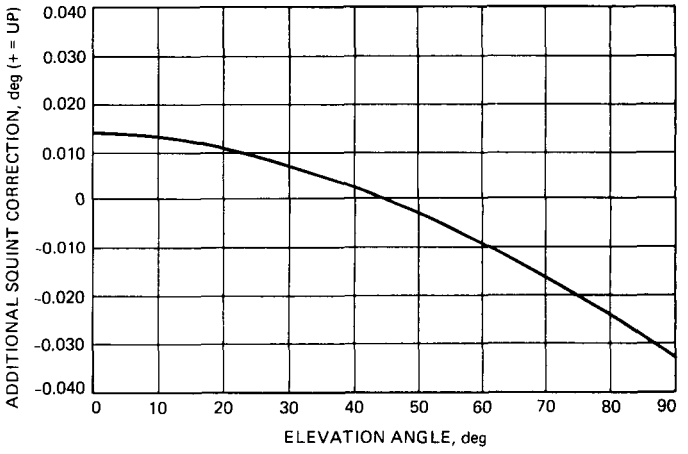


Fig. 8. 70-m antenna additional squint correction needed versus elevation angle.

Pointing a Ground Antenna at a Spinning Spacecraft Using Conscan–Simulation Results

A. Mileant and T. Peng
Telecommunications Systems Section

This article presents the results of an investigation of ground antenna pointing errors which are caused by fluctuations of the receiver AGC signal due to thermal noise and a spinning spacecraft. Transient responses and steady-state errors and losses are estimated using models of the digital Conscan (conical scan) loop, the FFT, and antenna characteristics. Simulation results are given for the on-going Voyager mission and for the upcoming Galileo mission, which includes a spinning spacecraft. The simulation predicts a 1-sigma pointing error of 0.5 to 2.0 mdeg for Voyager, assuming an AGC loop SNR of 35 to 30 dB with a scan period varying from 128 to 32 sec, respectively. This prediction is in agreement with the DSS 14 antenna Conscan performance of 1.7 mdeg for 32-sec scans as reported in earlier studies. The simulation for Galileo predicts 1-mdeg error with a 128-sec scan and 4-mdeg with a 32-sec scan under similar AGC conditions.

I. Introduction

In order to reduce the pointing error of the DSN ground antennas, a technique called Conscan has been successfully used for many years. In Conscan, angle tracking is accomplished by scanning the antenna around boresight in a circular pattern with constant angular offset, called the scan radius. The basic theory of Conscan is given in [1]. In the present implementation of the Conscan technique, the downlink signal is processed by a Fast Fourier Transform (FFT) algorithm.

In this article, the impact of downlink AGC fluctuations due to thermal noise and a spinning spacecraft on the pointing error of a ground antenna is investigated. In Section II the Conscan process is modeled as a digital phase-locked loop combined with the FFT algorithm. This analysis is subdivided into the following sections:

A. Modeling the downlink signal

B. Algorithm for estimating the pointing error

C. Impact of spacecraft motion on the pointing error

D. Impact of AGC SNR on the pointing error

E. Conscan closed-loop model

F. Pointing jitter and pointing loss

G. Transient response

Section III presents a computer simulation of the Conscan model. Predicted performance in terms of steady-state and transient responses, as well as time constants, are given as a function of loop gain, scan period, and signal-to-noise ratio. Specific topics discussed are:

A. Simulation model

B. Simulation results—choice of loop gain and scan period

C. Simulation versus actual performance for Voyager

D. Predicted performance for Galileo

E. Simulation results versus predicted performance

II. Analysis

A. Modeling of the Downlink Signal

The Conscan technique uses AGC samples of the ground receiver in order to estimate the ground antenna pointing angle. In general, the AGC samples are perturbed by several causes: by deliberate ground antenna scanning about its boresight, by thermal noise in the AGC loop, by spacecraft antenna mispointing, and by other factors such as changes in gain and weather conditions.

In our analysis, we first will assume that the AGC loop operates at very high SNR so that the AGC fluctuations due to thermal noise can be neglected. The effect of thermal noise will be addressed later in this analysis.

Let P_c be the average carrier power reaching the ground receiver when the receiving antenna is perfectly pointed and let $\beta(t)$ be the instantaneous pointing offset of the receiving antenna. Then, with imperfect ground antenna pointing, the signal power reaching the ground receiver will be

$$r(t) = P_c + l_r[\beta(t)] \qquad \text{(dBm)} \qquad (1)$$

Where $l_r(\cdot)$ represents the power loss due to pointing offset of the receiving antenna (a negative number in dB).

When a spinning spacecraft is tracked, $r(t)$ will experience periodic fluctuations of the form

$$\sum_{i=1}^{M} K_i \cos(\omega_i t + \phi_i) \qquad (2)$$

where K_i represents the maximum power deviation about the mean (in units of dB) at the frequency ω_i (which will be some multiple of the spacecraft's spin rate). Both K_i and ω_i are determined by the spacecraft's antenna gain pattern and dynamics.

The instantaneous pointing offset of the ground antenna, according to [1], can be expressed by the following equation

$$\beta(t) = \left[R^2 + \theta_e^2 + \theta_x^2 + 2R(\theta_x \cos \omega_T t + \theta_e \sin \omega_T t)\right]^{1/2} \qquad (3)$$

where

R = scan radius

θ_e = pointing error in elevation angle

θ_x = pointing error in cross-elevation angle

$\omega_T = 2\pi/T$, where T is the Conscan period

For small pointing offsets, the pointing loss due to *ground* antenna offset can be approximated by the least square fitting of a parabola to the antenna gain pattern. In our analysis we let

$$l_r(\beta) = K_r \beta^2 \qquad \text{(dB)} \qquad (4)$$

where K_r is a constant.

Combining the effects of scanning the ground antenna and a spinning spacecraft, the downlink AGC signal will be of the form

$$r(t) = P_c + l_r[\beta(t)] + \sum K_i \cos(\omega_i t + \phi_i) \qquad (5)$$

In Conscan implementation, the received AGC signal power, $r(t)$, is sampled every I seconds and stored in a one-dimensional array. Inserting Eqs. (3) and (4) into Eq. (5) and making $T = IN$ and $t = Ij$, we obtain the downlink AGC signal at the *sampling instants*, namely,

$$r_j = \left[P_c + K_r(R^2 + \theta_x^2 + \theta_e^2)\right] + 2K_r R\left[\theta_x \cos\left(\frac{2\pi j}{N}\right) + \theta_e \sin\left(\frac{2\pi j}{N}\right)\right] + \left[\sum K_i \cos(\omega_i Ij + \phi_i)\right] \qquad (6)$$

where the subscript j refers to the jth sample of a scan cycle, $j = 0, \ldots, N-1$. Subscript i refers to the ith contribution to the signal. The variables ϕ_i are random phases generated by a spinning spacecraft (for example, by its wobble and nutation). In general the spacecraft's spin rate is faster than the scan rate and should not be a multiple of the later (otherwise the scan rate must be changed accordingly). In this case, the phases ϕ_i will be approximately uniformly distributed in the $\{0, 2\pi\}$ interval. P_c as well as K_i ($i = 1, \ldots, M$), R, θ_x, and θ_e are

assumed to be constant during many scan periods. Equation (6) can be written as follows

$$r_j = P + c_j + n_j \tag{7}$$

where P represents the terms contained in the first pair of square brackets of Eq. (6), which correspond to the dc component; c_j represents the terms contained in the second pair of square brackets, which include the signal variation produced by scanning the ground antenna in a circular pattern, where θ_x and θ_e are the pointing errors that we want to correct; and n_j represents the terms inside the third pair of square brackets, which are the signal fluctuations produced by the spinning spacecraft.

B. Algorithm for Estimating the Pointing Error

In the present implementation of the Conscan technique, at the end of a scan period, N AGC samples are Fourier transformed with an FFT algorithm. The antenna pointing errors are estimated from the *first component* of the FFT (see Fig. 1). Since the FFT is just a fast implementation of the Discrete Fourier Transform, we will apply the theory of DFT to estimate the impact of a spinning spacecraft on the Conscan signal.

Let D^k be the kth component of the Discrete Fourier Transform operator defined by

$$D^k(r) = \left(\frac{1}{N}\right)\sum_{j=0}^{N-1} r_j\left[\cos\left(\frac{2\pi kj}{N}\right) + i\sin\left(\frac{2\pi kj}{N}\right)\right] \tag{8}$$

We rewrite this equation as follows

$$D^k(r) \doteq R(k) = R_R(k) + iR_I(k) \tag{9}$$

where $R_R(k)$ and $R_I(k)$ are the real and imaginary parts of the kth component of the DFT, $k = 0, 1, \ldots, N-1$. Carrying out the above DFT operation on Eq. (6) we obtain

Real part

$$R_R(k) = \begin{cases} P + N_R(0) & \text{for } k = 0 \\ K_r R\theta_x + N_R(1) & \text{for } k = 1 \\ N_R(k) & \text{otherwise} \end{cases} \tag{10}$$

Imaginary part

$$R_I(k) = \begin{cases} N_I(0) = 0 & \text{for } k = 0 \\ K_r R\theta_e + N_I(1) & \text{for } k = 1 \\ N_I(k) & \text{otherwise} \end{cases} \tag{11}$$

where P equals the sum of terms in the first pair of square brackets in Eq. (6), whose value is of no interest in our analysis. As we see from Eqs. (10) and (11), θ_x and θ_e can be estimated from $R_R(1)$ and $R_I(1)$, namely,

$$\widehat{\theta}_x = \frac{R_R(1)}{K_r R}$$

$$\widehat{\theta}_e = \frac{R_I(1)}{K_r R} \tag{12}$$

$N_R(1)$ and $N_I(1)$ result from the modulation of the downlink signal produced by a spinning spacecraft. Since the above terms superimpose to the terms containing θ_x and θ_e (see Eqs. 10 and 11), they have the effect of an additive noise which corrupts the estimation of the ground antenna pointing error. In what follows, the impact of $N_R(1)$ and $N_I(1)$ on $\widehat{\theta}_x$ and $\widehat{\theta}_e$ will be investigated.

C. Impact of a Spinning Spacecraft on the Pointing Error

We begin by taking the DFT on n_j, the terms inside the third pair of square brackets in Eq. (6), namely

$$D^k(n) = D^k\left[\sum K_i \cos(\omega_i Ij + \phi_i)\right]$$

$$\doteq N(k) = N_R(k) + iN_I(k) \tag{13}$$

where $N_R(k)$ and $iN_I(k)$ are the real and imaginary parts of $N(k)$. Using Eqs. (A-21) and (A-23) of the Appendix we obtain

$$N_R(1) = \left(\frac{1}{2N}\right)\sum_{\text{all } i} K_i \sin\gamma_{0i}$$

$$\times\left[\cos\phi_i(C_{1i} + C_{2i}) - \sin\phi_i(C_{3i} + C_{4i})\right]$$

$$N_I(1) = \left(\frac{1}{2N}\right)\sum_{\text{all } i} K_i \sin\gamma_{0i} \tag{14}$$

$$\times\left[\cos\phi_i(C_{3i} - C_{4i}) - \sin\phi_i(-C_{1i} + C_{2i})\right]$$

where

$$C_{1i} = \frac{\cos\gamma_{1i}}{\sin\gamma_{3i}} \qquad C_{2i} = \frac{\cos\gamma_{2i}}{\sin\gamma_{4i}}$$

$$C_{3i} = \frac{\sin\gamma_{1i}}{\sin\gamma_{3i}} \qquad C_{4i} = \frac{\sin\gamma_{2i}}{\sin\gamma_{4i}}$$

$$\gamma_{0i} = \frac{\pi IN}{T_i}$$

$$\gamma_{1i} = \pi\left[\frac{I(N-1)}{T_i} + \frac{1}{N}\right] \qquad \gamma_{2i} = \pi\left[\frac{I(N-1)}{T_i} - \frac{1}{N}\right]$$

$$\gamma_{3i} = \pi\left[\frac{I}{T_i} - \frac{1}{N}\right] \qquad \gamma_{4i} = \pi\left[\frac{I}{T_i} + \frac{1}{N}\right] \tag{15}$$

When ϕ_i are uniformly distributed, the evaluation of $N_R(1)$ and $N_I(1)$ is straightforward. It is shown in the Appendix that $N_R(1)$ and $N_I(1)$ are processes with zero mean, and variances (expressed in dB^2) of

$$\text{var}\{N_R(1)\} \doteq \sigma_{Rs}^2 = \left(\frac{1}{8N^2}\right)\sum_{\text{all } i}(K_i \sin\gamma_{0i})^2 \times \left[(C_{1i}+C_{2i})^2 + (C_{3i}+C_{4i})^2\right]$$

$$\text{var}\{N_I(1)\} \doteq \sigma_{Is}^2 = \left(\frac{1}{8N^2}\right)\sum_{\text{all } i}(K_i \sin\gamma_{0i})^2 \times \left[(C_{3i}-C_{4i})^2 + (-C_{1i}+C_{2i})^2\right] \tag{16}$$

In order to gain more insight into the operation of the Conscan algorithm, we can think of the FFT as digital filtering. In this context, we can compute the contribution to the variances σ_{Rs}^2 and σ_{Is}^2 in terms of the ith modulation component of a spinning spacecraft and the FFT's transfer function, namely

$$\sigma_{Rsi}^2 = F_R^2(i)(K_i \sin\gamma_{0i})^2$$

$$\sigma_{Isi}^2 = F_I^2(i)(K_i \sin\gamma_{0i})^2 \tag{17}$$

where the transfer functions for the real and imaginary parts are obtained from Eq. (16), namely

$$F_R^2(i) = \left(\frac{1}{8N^2}\right)\left[(C_{1i}+C_{2i})^2 + (C_{3i}+C_{4i})^2\right]$$

$$F_I^2(i) = \left(\frac{1}{8N^2}\right)\left[(C_{3i}-C_{4i})^2 + (-C_{1i}+C_{2i})^2\right] \tag{18}$$

Figure 2 shows the frequency response of $F^2(f) = F_R^2(f) + F_I^2(f)$, the magnitude squared of the first FFT's component. Note that the transfer function of the FFT processing is periodic. The period equals I, which is the AGC sampling time (usually 1 sec). So, the frequencies f_i ($f_i = \omega_i/2\pi$), which are multiples of $1/I$, hurt the Conscan estimator the most. Minima of $F^2(f)$ are 23 dB below the maximum. Note that $F^2(f)$ has nulls at multiples of $1/T$. Figure 2 also shows the ratio of $F_R^2(f)/F_I^2(f)$ versus frequency.

With this information about the properties of the transfer function of the FFT algorithm, we can select the Conscan period T so as to minimize the impact of unwanted frequencies, ω_i. Ideally, we would like to have freedom in selecting both the AGC sampling time I, and the Conscan period T, so that the unwanted frequencies will fall at the nulls or where the FFT's frequency response is minimal. For example, we would like to select the sampling time I so that the following condition is met

$$\frac{\omega_1}{2\pi} \approx \left(\frac{n}{I} + 0.5\right)$$

where ω_1 is the most significant component of the spin rate and n is an integer.

D. Effect of AGC SNR on the Pointing Error

Again let P_c be the nominal carrier power reaching the ground receiver and SNR be the signal-to-noise ratio in the AGC loop. Then the noise variance in the AGC loop will be $\sigma_N^2 = P_c/SNR$. The instantaneous *signal plus thermal noise* power of the jth AGC sample will be

$$r_j = 20.0 \log\left(\sqrt{P_c} + V_{Nj}\right) \tag{19}$$

where V_{Nj} is a zero mean Gaussian random variable with variance σ_N^2. The standard deviation of the r_j sample (assuming that only thermal noise is present) will be approximately

$$\sigma_r \approx 20.0 \log \left[\frac{\left(\sqrt{P_c} + \sigma_N\right)}{\sqrt{P_c}} \right] \quad \text{(dB)} \qquad (20)$$

It is shown in the Appendix that the variances of the real and imaginary parts at the output of the FFT in terms of the variance of the AGC thermal noise are

$$\sigma_{Rt}^2 = \sigma_{It}^2 = \frac{\sigma_r^2}{2N} \quad (\text{dB}^2) \qquad (21)$$

The overall variances at the output of the FFT will be simply the sum of the individual variances due to spacecraft spin and receiver thermal noise, namely

$$\begin{aligned} \sigma_R^2 &= \sigma_{Rs}^2 + \sigma_{Rt}^2 \\ \sigma_I^2 &= \sigma_{Is}^2 + \sigma_{It}^2 \end{aligned} \qquad (22)$$

where σ_{Rs}^2 and σ_{Is}^2 are given by Eq. (16) and σ_{Rt}^2 and σ_{It}^2 by Eq. (21).

E. Conscan Closed-Loop Model

So far we have discussed the open-loop estimation of the pointing errors. In order to proceed with this analysis, we define the *closed-loop pointing errors* in cross-elevation and elevation as follows

$$\begin{aligned} \phi_{x(n+1)} &\doteq \theta_{xn} - \hat{\theta}_{xn} \\ \phi_{e(n+1)} &\doteq \theta_{en} - \hat{\theta}_{en} \end{aligned} \qquad (23)$$

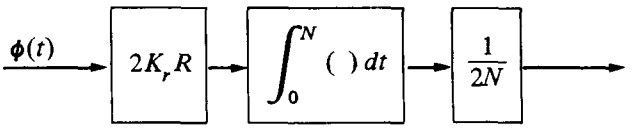
where $\hat{\theta}_x$ and $\hat{\theta}_e$ are the Conscan estimates of the pointing offsets θ_x and θ_e, respectively. We will treat all of the above angles as *continuous* variables of time and model the DFT as an analog multiplication and integration. This approach is allowable because the AGC sampling time I is much smaller than the Conscan update time, $T = NI$. Being in closed-loop, we substitute ϕ_x for θ_x and ϕ_e for θ_e in the third term of Eq. (6) and rewrite that term in vector notation as follows:

$$2K_r R \left[\cos \frac{2\pi t}{N} \quad \sin \frac{2\pi t}{N} \right] \begin{bmatrix} \phi_x(t) \\ \phi_e(t) \end{bmatrix} \qquad (24)$$

The DFT algorithm (modeled here as an analog operation) multiplies this input by the vector

$$\left[\cos \frac{2\pi t}{N} \quad i \sin \frac{2\pi t}{N} \right]$$

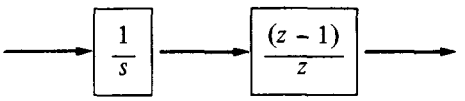
integrates from 0 to N, and divides the result by N. This operation is represented schematically below.



where

$$\phi(t)^T = [\phi_x(t) \quad \phi_e(t)] \qquad (25)$$

The factor of 1/2 results from the multiplication of a cosine by a cosine and a sine by a sine. Double frequency terms are integrated to zero. Thus, the DFT operation for a large N can be modeled as an integrate-and-dump device with the following transfer function in the hybrid s/z-domain.



In the feedback path of the Conscan loop, the estimated errors in antenna pointing are scaled down by a factor G, and the antenna pointing corrections, $\hat{\theta}_x$ and $\hat{\theta}_c$, are obtained. This corresponds to the following set of difference equations

$$\begin{aligned} \hat{\theta}_{xn} &= \hat{\theta}_{x(n-1)} + G\hat{\phi}_{xn} \\ \hat{\theta}_{en} &= \hat{\theta}_{e(n-1)} + G\hat{\phi}_{en} \end{aligned} \qquad (26)$$

where $\hat{\phi}_x$ and $\hat{\phi}_e$ are the estimates of ϕ_x and ϕ_e, respectively. The subscript n indicates the nth scan period. In Fig. 3 the above difference equation has the following z-domain transfer function (summer)

$$S(z) \doteq \frac{\hat{\theta}(z)}{\hat{\phi}(z)} = \frac{z}{(z-1)} \qquad (27)$$

Because $\hat{\theta}_x$ and $\hat{\theta}_e$ are modeled as continuous variables of time, we need to convert the discrete variables $\hat{\theta}_x(z)$ and $\hat{\theta}_e(z)$ to their continuous counterparts, namely, we need a Digital-to-

Analog converter (D/A) in the feedback path of our loop model. The transfer function for the D/A is

$$\frac{(1-e^{-sN})}{s} \tag{28}$$

where $e^{-sN} \doteq z^{-1}$.

By combining the above elements into a block diagram, Fig. 3 is obtained. Now the above hybrid s/z-loop model may be converted to a z-domain model. Neglecting for the moment the noise term $N(z)$, we see by inspection of Fig. 3 that

$$\boldsymbol{X}'(s) = \frac{2K_r R\boldsymbol{\theta}(s)}{s} - \boldsymbol{X}'(z)\,\frac{\left[\frac{(z-1)}{z}\right]G}{(Ns^2)} \tag{29}$$

Taking the z-transform of Eq. (29) and simplifying, we obtain

$$\boldsymbol{X}'(z) = 2K_r R\left[\frac{\boldsymbol{\theta}(s)}{s}\right]^* - \boldsymbol{X}'(z)\,\frac{G}{(z-1)} \tag{30}$$

where the asterisk denotes the z-transform operation.

With this transformation, Fig. 3 has the equivalent block diagram of Fig. 4, which is entirely in the z-plane. From Fig. 4 we see that the overall Conscan open loop transfer function is

$$G(z) = \frac{Y(z)}{X(z)} = \frac{G}{(z-1)} \tag{31}$$

and the closed-loop transfer function is

$$H(z) = \frac{Y(z)}{\left[\frac{\theta(s)}{s}\right]^*} = \frac{G(z)}{[1+G(z)]} = \frac{G}{[z+(G-1)]} \tag{32}$$

Having obtained $H(z)$, we now need the following integral

$$I_1 = \frac{1}{i2\pi}\oint H(z)H(z^{-1})\frac{dz}{z} \tag{33}$$

Using Table III of [2], it is found that

$$I_1 = \frac{G}{(2-G)} \tag{34}$$

Assuming then the noise sample $N_{n+1}(1)$ is independent of $N_n(1)$ for all n, it can be shown that the steady-state closed-loop variances of the pointing error are

$$\sigma_x^2 = \frac{I_1\sigma_R^2}{(K_r R)^2} \qquad (\text{in units of } R^2)$$

$$\sigma_e^2 = \frac{I_1\sigma_I^2}{(K_r R)^2} \qquad (\text{in units of } R^2) \tag{35}$$

where σ_R^2 and σ_I^2 are the open-loop variances at the output of the FFT and are given by Eq. (22).

Before moving to the next topic it should be noted that the transfer function $H(z)$ of the Conscan loop, Eq. (32), has a single pole at $z = 1 - G$. The stability criterion requires that $|1 - G| < 1$ in order for the pole to remain inside the unit circle. This puts an upper bound on the loop gain which has to be less than 2. Control theory predicts that the Conscan loop will have a bounded steady-state error to a ramp or velocity input. In practice, the velocity components of the pointing axis are compensated by predicts, while Conscan compensates for step pointing errors.

F. Pointing Jitter and Pointing Loss

Equation (4) relates the ground antenna pointing loss to the pointing error. In order to estimate the closed-loop pointing loss, we substitute ϕ_x for θ_x and ϕ_e for θ_e in Eq. (3). Then we insert Eq. (3) into Eq. (4) and integrate over one scan period to obtain

$$1_r = K_r\left(R^2 + \phi_x^2 + \phi_e^2\right) \tag{36}$$

where ϕ_x and ϕ_e are random variables with zero mean (assuming a step input) and the variance is given by Eq. (35). Invoking the central limit theorem, we can assume that the closed-loop pointing errors ϕ_x and ϕ_e have a Gaussian distribution. Being derived from quadrature processes, they are mutually independent. With the above assumptions, we take the expected value of Eq. (36) and obtain the closed-loop pointing loss, L, namely

$$L \doteq \mathbf{E}\{1_r\}$$

$$= K_r\left[R^2 + \mathbf{E}(\phi_x^2) + \mathbf{E}(\phi_e^2)\right] \tag{37}$$

$$= K_r(R^2 + \sigma_x^2 + \sigma_e^2) \qquad (\text{dB}) \tag{38}$$

G. Transient Response

Let $\theta_{xn} = \theta_{x1}$ and $\theta_{en} = \theta_{e1}$ for all n. Inserting Eq. (26) in Eq. (23) and taking the expected value, it can be shown that

$$\phi_{xn} = \theta_{x1} r^{(n-1)}$$
$$\phi_{en} = \theta_{e1} r^{(n-1)} \tag{39}$$

or, in vector notation

$$\boldsymbol{\phi}_n = \boldsymbol{\theta}_1 r^{(n-1)} \tag{40}$$

where

$$r \doteq 1 - G \qquad (0 < r < 1)$$

We now define the time constant, τ, as the time in seconds that it takes for pointing error $\boldsymbol{\phi}_n$ to decay to the value of $\boldsymbol{\theta}_1/e$, where e is the base of the natural logarithm. Solving for the number of Conscan periods, $n-1$, necessary for θ_1 to decrease by $1/e$ and for the corresponding τ, it can be shown that

$$n - 1 = -\frac{1}{\ln(r)} \tag{41}$$

and

$$\tau = -\frac{T}{\ln(r)} \tag{42}$$

Figure 5 depicts typical τ values for DSN antennas.

III. Computer Simulation

A. Simulation Model

A computer program named CONSCAN.FOR using an FFT subroutine was written in order to check the above analysis. In our simulation, values typical for the Voyager and the Galileo missions were used.

In the actual implementation of the Conscan algorithm at the DSN stations the FFT computation can start at any time inside a Conscan period. In order to account in our simulation for this time shift between the FFT reference point and AGC samples, the estimates of the pointing error were computed with the following modification to Eq. (12)

$$\text{MAG} = \frac{\left[R_R^2(1) + R_I^2(1)\right]^{1/2}}{(K_r R)} \tag{43}$$

$$\text{ANG} = \tan^{-1}\left[\frac{R_I(1)}{R_R(1)}\right] + \gamma \tag{44}$$

where γ represents the relative angle offset between the AGC table of the antenna scan cycle. In Eq. (44), a four-quadrant arc tangent is computed. Finally, the estimates of the pointing error are computed

$$\theta_x = \text{MAG} \cdot \cos(\text{ANG})$$
$$\theta_e = \text{MAG} \cdot \sin(\text{ANG}) \tag{45}$$

This revised estimation algorithm for the pointing error shows a slight crosscoupling between θ_x and θ_e. In this simulation, in order to mimic more closely an actual antenna, the antenna pointing corrections were done gradually in an interval of eight to ten samples with a slight overshoot before the final point.

Figures 6 through 9 display the results of the simulation. The variances of the pointing error were averaged over 550 Conscan periods. The computer-simulated results agree very closely with the equations derived in this analysis.

The following values were used in the simulation: $K_r R^2$ = 0.1 dB for both the 64- and 70-m antennas; scan period T = 32, 64, and 128 sec; and sampling time I = 1 sec. Conscan loop gain G = 0.05, 0.3, and 0.6. P_c = −145 dBm, and the AGC loop SNR was between 20 and 50 dB.

For the Galileo spacecraft, the effect of the spin on the transmitted signal was modeled as follows (see Eq. 2)

$$\Sigma K_i \cos(\omega_i t + \phi_i) = K_t \alpha^2 \qquad \text{(dB)} \tag{46}$$

where α, the effective offset from the correct pointing of the spacecraft antenna, is defined as

$$\alpha(t) = \Big[\alpha_1^2 + \alpha_2^2 + \alpha_3^2 + 2\alpha_1\alpha_2\cos(\omega_1 t + \phi_1) + 2\alpha_1\alpha_3\cos(\omega_2 t + \phi_2) + 2\alpha_2\alpha_3\cos(\omega_3 t + \phi_3)\Big]^{1/2} \tag{47}$$

Here

α_1 = combined pointing error due to Earth-fitting error, spacecraft/Earth drift, and attitude determination error

α_2 = pointing error due to nutation

α_3 = pointing error due to wobble, mechanical and electrical misalignments

In our simulation we made $K_1 = 2\alpha_1\alpha_2 = 1$ dB, $K_2 = 2\alpha_1\alpha_3 = 0.2$ dB, and $K_3 = 2\alpha_2\alpha_3 = 0.2$ dB. Other variables are defined as follows: ω_1 = nutation spin rate, ω_2 = wobble spin rate, $\omega_3 \doteq \omega_1 - \omega_2$. The following values were assumed: wobble period $T_1 = 19.048$ sec (nominal or low spin rate); nutation period $T_2 = 14.583$ sec ($T_2 \approx T_1/1.3$); $K_t = -77.5$ dB/deg^2 (curve fit to Galileo X-band antenna gain pattern).

B. Simulation Results—Choice of Loop Gain and Scan Period

From Figs. 6 through 9 we can make the following observations:

Increasing the gain G decreases the time constant and, hence, the lock-up time, although it increases the pointing fluctuations and pointing loss (Figs. 6 and 7). This fact suggests the following operational strategy: choose a large gain value during lock-up, and a small gain value during tracking. This strategy has been successfully used with the Real-Time Combiner of the DSN Baseband Assembly.

Increasing the scan period decreases the pointing jitter (Figs. 8 and 9). This effect is especially pronounced for Galileo which has larger pointing jitter due to spacecraft spin. Increasing the scan period, however, also slows down the pointing correction process.

C. Simulation Versus Actual Performance for Voyager

The simulation result is comparable with the observed antenna Conscan performance in supporting Voyager. Figure 8 predicts a 1-sigma pointing error of 0.5 (128-sec scan) to 1.5 mdeg (32-sec scan) for Voyager, assuming thermal noise 35 dB below the carrier level in a 1-Hz AGC loop. In [3] the Conscan performance of the DSS 14 64-m antenna in pointing Voyager over 5 days in 1987 is reported (see Fig. A5 in [3]), with statistics given in Fig. A6 for individual scans with a 32-sec period. The standard deviation reported is 1.7 mdeg, in good agreement with our simulation result.

D. Predicted Performance for Galileo

In the case of Galileo, the Conscan performance is affected by both thermal noise in the receiver and the wobble and nutation reflected in the spacecraft signal. Figure 9 indicates that the thermal noise is the dominant effect when the AGC loop SNR is lower than about 30 dB. But the wobble and nutation effect dominates when the AGC is higher than 30 dB. As a result, the pointing error can be reduced by increasing the AGC SNR below 30 dB, but not beyond 30 dB.

It should be noted that the Conscan period must be chosen carefully in order to avoid harmonic relationship with the wobble and nutation processes. In case both wobble and nutation are present, the scan period of 64 sec should be avoided because its frequency, 0.0156 Hz, is very close to the difference between the wobble and nutation frequencies, 0.0162 Hz. Abnormal pointing errors were indeed observed in simulation.

E. Simulation Results Versus Predicted Performance

Simulation results give slightly more conservative values for the pointing error variances than the ones predicted by the closed-form solution, Eq. (35). The possible reasons are (a) in our simulation, the lock-up transients have partially effected the statistics of the pointing error; and (b) an error in the phase offset γ in Eq. (44) was deliberately introduced in our simulation in order to approximate the actual pointing imperfections.

The discrepancy can be illustrated with two cases. For Voyager, with $T = 32$ sec, $SNR = 35$ dB and $G = 0.6$, the standard deviation of the pointing error is 0.6 mdeg according to the closed-form solution and 1 mdeg according to the simulation. For Galileo, with the same conditions, the standard deviation of the pointing error is 2 mdeg according to the closed-form solution and 3 mdeg according to the simulation.

IV. Conclusion

A model has been developed to analyze and simulate the performance of DSN antenna Conscan as affected by the fluctuations in the receiver AGC signal. The effects of the receiver thermal noise and of the spinning spacecraft were analyzed and simulated. The simulation results agreed well with the observed performance of the DSS 14 antenna supporting Voyager. The simulation results show a standard deviation of 1.5 mdeg at X-band with 32-sec scan for a 35-dB AGC SNR. Observed performance at DSS 14 [3] for Voyager under similar AGC SNR conditions was about 1.7 mdeg for individual scans.

The simulation results suggested that the Conscan loop gain and the scan period can be chosen to optimize the pointing performance. A higher gain and a shorter scan period can reduce pointing acquisition time at the beginning of the track. A lower gain and a longer scan period can reduce the standard deviation of pointing jitters, and, therefore, the loss, during track.

The simulation result for Galileo indicated higher pointing jitter than Voyager because of the additional effect of spacecraft wobble and nutation on the ground receiver AGC signal. The spacecraft effect is dominated by the ground receiver thermal noise effect for the AGC SNR below 30 dB and becomes dominant when the SNR is above 30 dB. Figure 9 predicts a relative constant pointing error in the latter region. This region is typical for Galileo support when the carrier signal level is at least −155 dBm.

The Conscan pointing error for Galileo under such conditions was between 1 mdeg (128-sec scan) and 4 mdeg (32-sec scan). The corresponding losses versus the standard deviations of pointing errors are given in Fig. 10.

It was noted that the choice of a Conscan period for Galileo must avoid harmonic relationships with the Galileo wobble and nutation frequencies. The results of the analysis indicated that the 64-sec scan should not be used for Galileo when both wobble and nutation are present.

Acknowledgment

The authors are thankful to Timothy T. Pham of the Telecommunications Systems Section for his assistance with computer simulation and plotting, and to Michael Wert of the TDA Mission Support and DSN Operations Section for his comment.

References

[1] J. E. Ohlson and M. S. Reid, *Conical-Scan Tracking with the 64-m Diameter Antenna at Goldstone*, JPL Technical Report 32-1605, Jet Propulsion Laboratory, Pasadena, California, October 23, 1976.

[2] E. I. Jury, *Theory and Applications of the Z-Transform Method*, Malabar, Florida: R. E. Krieger Publishing Co., 1982.

[3] C. N. Guiar et al., "DSS 14 Antenna Calibration for GSSR/VLA Saturn Radar Experiments," *TDA Progress Report 42-93*, vol. January–March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 309–337, May 15, 1988.

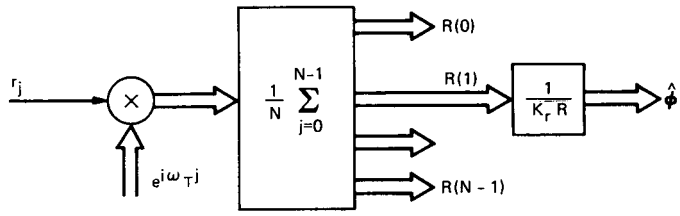


Fig. 1. Open-loop estimation of the pointing errors.

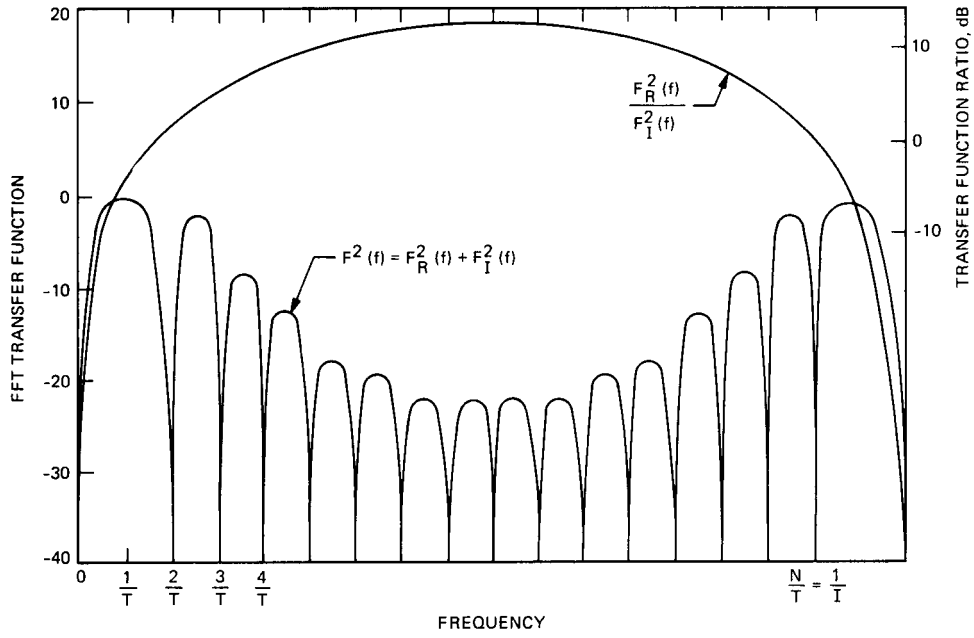


Fig. 2. Total (Real + Imaginary) FFT Transfer Function and Transfer Function Ratio (Real/Imaginary).

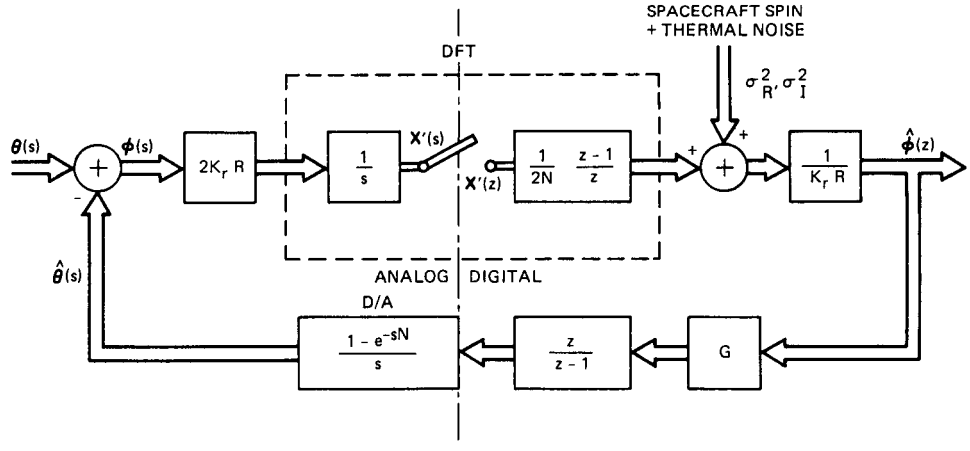


Fig. 3. Closed-loop hybrid s/z-domain model of the Conscan process.

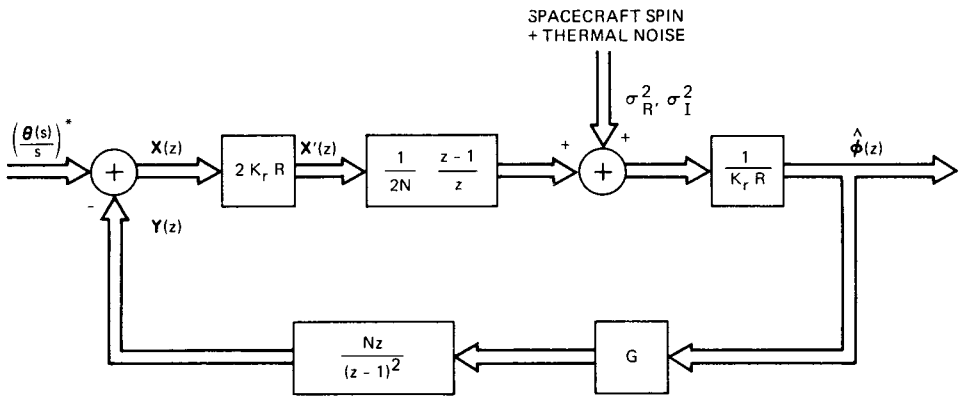


Fig. 4. *Z*-domain model of the Conscan process.

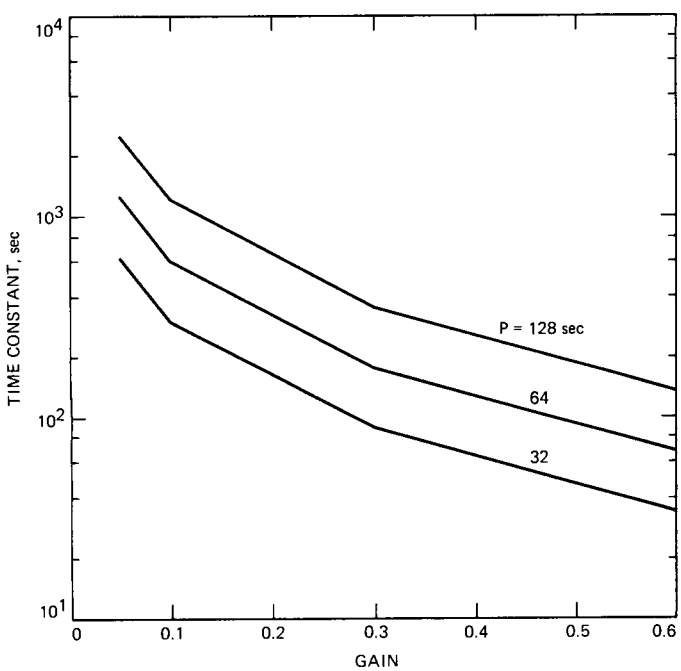


Fig. 5. Time constant versus closed-loop gain.

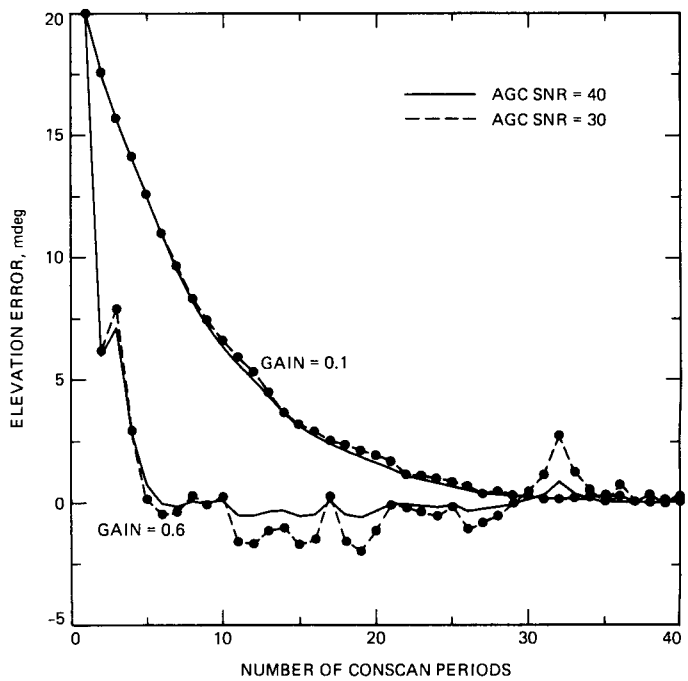


Fig. 6. Transient response of pointing error simulation for Voyager spacecraft (a typical case).

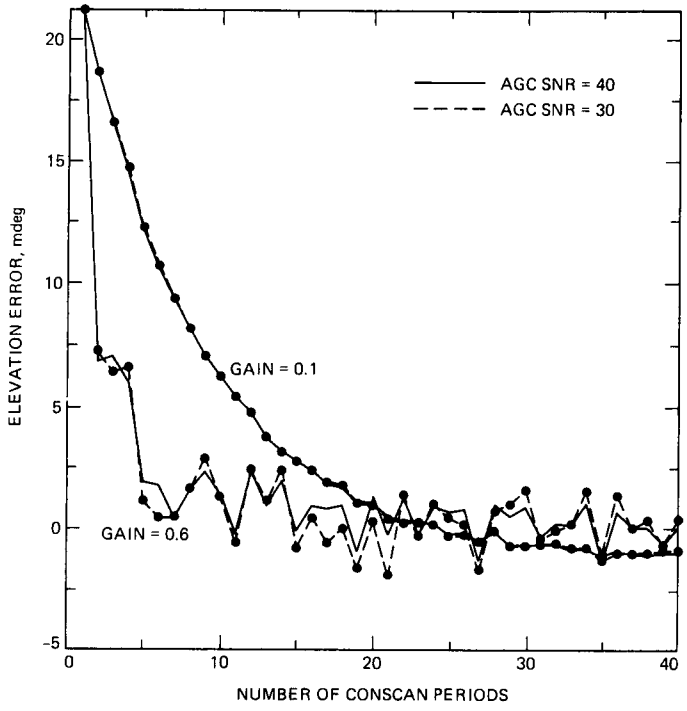


Fig. 7. Transient response of pointing error simulation for Galileo spacecraft (a typical case).

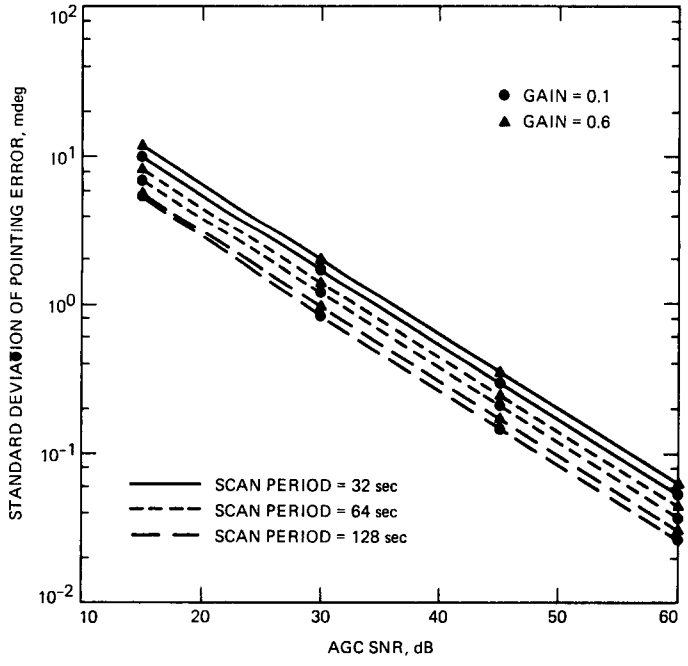


Fig. 8. Steady-state pointing error versus AGC SNR simulation for Voyager spacecraft.

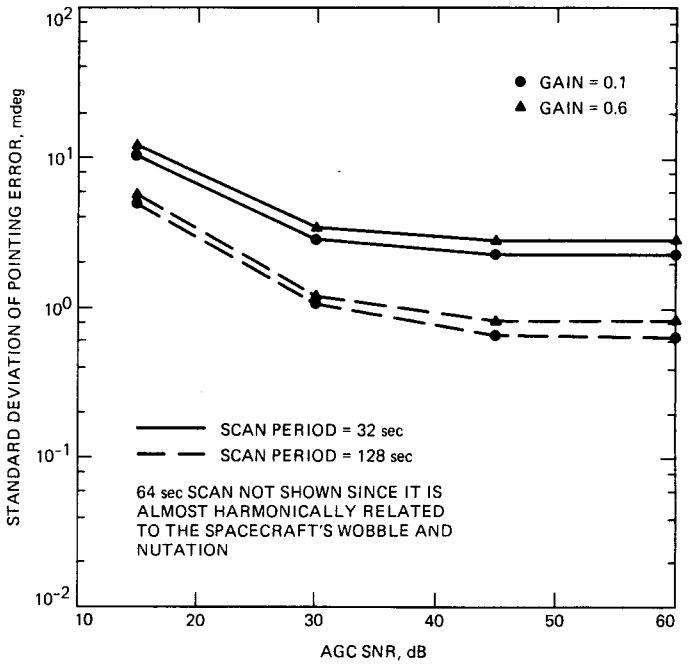


Fig. 9. Steady-state pointing error versus AGC SNR simulation for Galileo spacecraft.

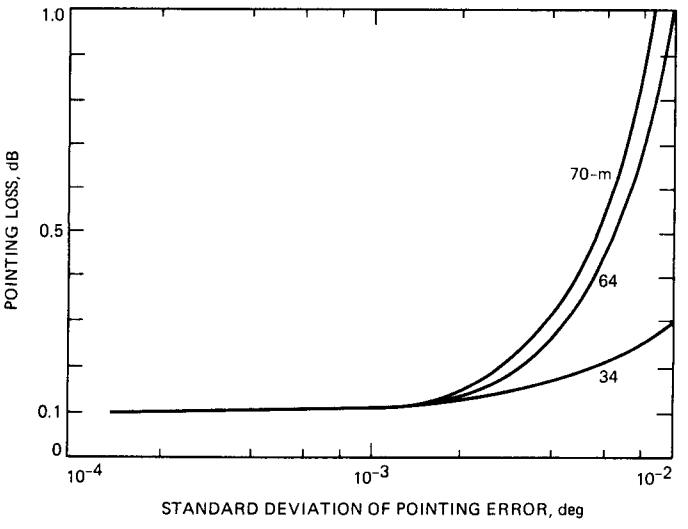


Fig. 10. Pointing loss versus standard deviation of the pointing error at X-band for 34-, 64-, and 70-m antennas. Nominal pointing loss due to Conscan = 0.1 dB.

Appendix A

Derivation of the Mean Value and Variance at the Output of DFT

Let D^k be the Discrete Fourier Transform operator defined as follows

$$D^k(S) = \left(\frac{1}{N}\right) \sum_{j=0}^{N-1} S_i \left[\cos\left(\frac{2\pi kj}{N}\right) + i \sin\left(\frac{2\pi kj}{N}\right)\right] \quad \text{(A-1)}$$

$$\doteq R(k) = R_R(k) + iR_I(k) \quad \text{(A-2)}$$

where $R_R(k)$ and $R_I(k)$ are the real and imaginary parts of $R(k)$, respectively.

We want to find the mean value and the variance of $R(k)$ for the following three cases

$$S_i = \begin{cases} S & \text{(A-3)} \\ S\cos\left(\dfrac{2\pi Ij}{T_i}\right) & \text{so that } K = \dfrac{NI}{T_i} \text{ is an integer} \quad \text{(A-4)} \\ S\cos\left(\dfrac{2\pi Ij}{T_i} + \phi_i\right) & \text{for any nonrandom } T_i \text{ and random } \phi_i, \text{ uniformly distributed} \quad \text{(A-5)} \end{cases}$$

where S is a random variable assumed to be constant during one Conscan period. And we define

$$\mathrm{E}\{S\} \doteq \mu_S \quad \text{and} \quad \mathrm{var}\{S\} \doteq \sigma_S^2 \quad \text{(A-6)}$$

$$\mathrm{E}\{D^k(r)\} \doteq \mu(k) = \mu_R(k) + i\mu_I(k) \quad \text{(A-7a)}$$

$$\mathrm{E}\{D^k(r)^2\} - \mu(k)^2 \doteq \sigma^2(k) = \sigma_R^2(k) + \sigma_I^2(k) \quad \text{(A-7b)}$$

A. $S_i = S$

Using Eq. (A-1) it is easy to see that

$$\mu_R(k) = \begin{cases} \mu_S & \text{for } k = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{(A-8)}$$

$$\mu_I(k) = 0 \quad \text{for all } k \quad \text{(A-9)}$$

$$\sigma_R^2(k) = \begin{cases} \sigma_S^2 & \text{for } k = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{(A-10)}$$

$$\sigma_I^2(k) = 0 \quad \text{for all } k \quad \text{(A-11)}$$

B. S_i Periodic as Defined by Eq. (A-4)

With $K = NI/T_i$ = integer, we have

$$S\cos\left(\frac{2\pi Ij}{T_i}\right) = S\cos\left(\frac{2\pi Kj}{N}\right) \quad \text{(A-12)}$$

Using Eq. (A-1) we obtain

$$\mu_R(k) = \begin{cases} \dfrac{\mu_S}{2} & \text{for } k = K \\ 0 & \text{otherwise} \end{cases} \quad \text{(A-13)}$$

$$\mu_I(k) = 0 \quad \text{for all } k \quad \text{(A-14)}$$

$$\sigma_R^2(k) = \begin{cases} \dfrac{\sigma_S^2}{4} & \text{for } k = K \\ 0 & \text{otherwise} \end{cases} \quad \text{(A-15)}$$

$$\sigma_I^2(k) = 0 \quad \text{for all } k \quad \text{(A-16)}$$

C. S_i Periodic With Any Period Length and Phase Shift

Let

$$S_i = S\cos\left(\frac{2\pi Ij}{T_i}\right) + \phi_i \quad \text{(A-17)}$$

$$= S\left[\cos\phi_i \cos\left(\frac{2\pi Ij}{T_i}\right) - \sin\phi_i \sin\left(\frac{2\pi Ij}{T_i}\right)\right] \quad \text{(A-18)}$$

Using Eq. (A-1) for the real part, we let $\omega_1 = 2\pi I/T_i$ and $\omega_2 = 2\pi k/N$. Then

$$R_R(k) = \left(\frac{S}{N}\right)\sum_{j=0}^{N-1}\left[\cos\phi_i \cos\omega_1 j - \sin\phi_i \sin\omega_1 j\right]\cos\omega_2 j$$

$$= \left(\frac{S}{2N}\right)\sum_{j=0}^{N-1}\left[\cos\phi_i \,\text{Real}\left\{e^{i(\omega_1-\omega_2)j} + e^{i(\omega_1+\omega_2)j}\right\}\right.$$

$$\left. - \sin\phi_i \,\text{Imag.}\left\{e^{i(\omega_1+\omega_2)j} + e^{i(\omega_1+\omega_2)j}\right\}\right]$$

$$= \left(\frac{S}{2N}\right)\left[\cos\phi_i \,\text{Real}\left\{\frac{1-e^{i(\omega_1-\omega_2)N}}{1-e^{i(\omega_1-\omega_2)}} + \frac{1-e^{i(\omega_1+\omega_2)N}}{1-e^{i(\omega_1+\omega_2)}}\right\}\right.$$

$$\left. - \sin\phi_i \,\text{Imag.}\left\{\frac{1-e^{i(\omega_1-\omega_2)N}}{1-e^{i(\omega_1-\omega_2)}} + \frac{1-e^{i(\omega_1+\omega_2)N}}{1-e^{i(\omega_1+\omega_2)}}\right\}\right] \quad \text{(A-19)}$$

But

$$e^{\pm i\omega_2 N} \equiv 1$$

and

$$\frac{1-e^{i\omega_1 N}}{1-e^{i(\omega_1 \pm \omega_2)}} = e^{i[(\omega_1(N-1)\mp\omega_2]}\,\frac{\sin\left(\frac{\omega_1 N}{2}\right)}{\sin\left[\frac{(\omega_1 \pm \omega_2)}{2}\right]} \quad \text{(A-20)}$$

Using Eq. (A-20) in Eq. (A-19) and simplifying we obtain

$$R_R(k) = \left(\frac{S}{2N}\right)\sin\gamma_0\left[\cos\phi_i\,(C_1 + C_2) - \sin\phi_i\,(C_3 + C_4)\right] \quad \text{(A-21)}$$

where

$$C_1 = \frac{\cos\gamma_1}{\sin\gamma_3} \qquad C_2 = \frac{\cos\gamma_2}{\sin\gamma_4}$$

$$C_3 = \frac{\sin\gamma_1}{\sin\gamma_3} \qquad C_4 = \frac{\sin\gamma_2}{\sin\gamma_4}$$

$$\gamma_0 = \frac{\pi IN}{T_i}$$

$$\gamma_1 = \pi\left[\frac{I(N-1)}{T_i} + \frac{k}{N}\right] \qquad \gamma_2 = \pi\left[\frac{I(N-1)}{T_i} - \frac{k}{N}\right]$$

$$\gamma_3 = \pi\left[\frac{I}{T_i} - \frac{k}{N}\right] \qquad \gamma_4 = \pi\left[\frac{I}{T_i} + \frac{k}{N}\right] \quad \text{(A-22)}$$

Note that C_i and γ_i $(i = 1, \ldots, 4)$ are functions of the variable k. Repeating the above steps for the imaginary part of $R(k)$ we obtain

$$R_I(k) = \left(\frac{S}{2N}\right)\sin\gamma_0\left[\cos\phi_i\,(C_3 - C_4) - \sin\phi_i\,(-C_1 + C_2)\right] \quad \text{(A-23)}$$

Assuming that ϕ_i is uniformly distributed in the interval $\{0, 2\pi\}$, we obtain

$$\mu_R = \mu_I = 0 \qquad \text{for all } k \quad \text{(A-24)}$$

$$\sigma_R^2(k) = \sigma_S^2\left(\frac{1}{8N^2}\right)(\sin\gamma_0)^2 \times\left[(C_1 + C_2)^2 + (C_3 + C_4)^2\right] \quad \text{(A-25)}$$

$$\sigma_I^2(k) = \sigma_S^2\left(\frac{1}{8N^2}\right)(\sin\gamma_0)^2 \times\left[(C_3 - C_4)^2 + (-C_1 + C_2)^2\right] \quad \text{(A-26)}$$

where again σ_S^2 is the variance of S and other variables are defined in Eq. (A-22).

D. White Noise

Assuming that the variance of each voltage sample due to thermal noise is σ_v^2, the variance of the sum of N samples will be $N\sigma_v^2$. The FFT algorithm performs the N complex sums and divides the result by N. Hence the total variance of the complex FFT output will be

$$\sigma_T^2 = \sigma_v^2 N\left(\frac{1}{N}\right)^2 = \frac{\sigma_v^2}{N} \tag{A-27}$$

Since in our implementation of the DFT the output is split into its real and imaginary parts, the variance of each of them will be one-half of σ_v^2.

C-2